RICE UNIVERSITY

Stochasticity and Cell Population Heterogeneity in an Artificial lac Operon Genetic Network

by

Michail Stamatakis

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE

Doctor of Philosophy

APPROVED, PROPOSAL COMMITTEE:

[Signature]
Kyriacos Zygourakis, A. J. Hartsook
Professor
Chemical and Biomolecular Engineering

[Signature]
Sibani Lisa Biswal, Assistant Professor
Chemical and Biomolecular Engineering

[Signature]
Kathleen S. Matthews, Stewart Memorial Professor
Biochemistry and Cell Biology

HOUSTON, TEXAS

MAY 2009
INFORMATION TO USERS

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleed-through, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.
Abstract

Stochasticity and Cell Population Heterogeneity

in an Artificial lac Operon Genetic Network

by

Michail Stamatakis

The purpose of this work is two-fold: (1) to develop a novel mathematical and computational framework that incorporates the major sources of cell population heterogeneity and (2) to use this framework to demonstrate the effect of stochasticity on cell population heterogeneity in an artificial lac operon genetic network.

During the past decades, several approaches have been used to model heterogeneity in bacterial cell populations, each approach focusing on different source(s) of heterogeneity. However, a holistic approach that integrates all the major sources into a generic framework is still lacking. In this work we present a mathematical and computational framework that describes single cells or cell populations and takes into account stochasticity in reaction, division and DNA duplication, all of which constitute sources of cell population heterogeneity. We subsequently use this framework to demonstrate how stochasticity generates complex behavior and phenotypic heterogeneity in the case of an artificial lac operon genetic network, characteristic of positive feedback regulation. Our results show that stochasticity can enhance phenotypic heterogeneity, create or destroy bistability, and result in noise-induced transitions between attracting vicinities. We also found that it is possible to predict population averages with carefully constructed single cell models.
Acknowledgements

I would like to thank my adviser Prof. Kyriacos Zygourakis for his help and support in the most critical period of my PhD study at Rice University, and the members of my committee, Prof. Kathleen Matthews and Prof. Lisa Biswal for their guidance and constructive input on my work.

Furthermore, I am grateful to Prof. Gábor Balázsi from the University of Texas M. D. Anderson Cancer Center and the members of his research group, Bhaskar, Dmitry and Rhys, for the wonderful collaboration we had together. I would also like to thank Prof. Andreas Boudouvis from the National Technical University of Athens for his advice, guidance and support during the tough times.

The research of Chapter 5 was performed under the advising of Dr. Nikos Mantzaris.

The time of my study in Houston would have been dull and monotonous had it not been for my great friends to whom I wholeheartedly extend my appreciation: Kostas S., Kostas B., Nikos, Maria, Venetia, Leda, Ilias, Arjun, Keith, Bibhash, Ryan, Abby, Stephanie, Clint and Jay (and any others whom I may be forgetting...)

Last but not least, I would like to thank my parents Dimitris and Alexia and my brother Tassos for supporting me in every step in my life. Without them nothing would have been possible...
# Table of Contents

Abstract ii
Acknowledgements iii
Table of Contents iv
List of Figures viii
List of Tables xv
Abbreviations xvi

Part I: Developing Mathematical and Computational Tools for Integrating the Major Sources of Cell Population Heterogeneity 1

Chapter 1: Introduction 2

1.1 The Study of Heterogeneity: Motivation 2
1.2 Historical Perspective 5
  1.2.1 Experimental Approaches 5
  1.2.2 Theoretical Approaches 13
1.3 Current Objectives 21

Chapter 2: Modeling the Major Sources of Stochasticity in Intracellular Processes at the Single Cell Level 24

2.1 Preliminary Definitions 24
2.2 Sources of Stochasticity in Intracellular Processes 25
  2.2.1 Chemical Reactions 26
  2.2.2 DNA Duplication 29
  2.2.3 Growth 30
  2.2.4 Division 30
2.3 The Single Cell M-Equation 32
2.4 Particular Expressions for the Propensity Functions 33
  2.4.1 Chemical Reactions 34
  2.4.2 Growth 35
  2.4.3 DNA Duplication 36
  2.4.4 Division 36
2.5 MC Simulation of the Single Cell M-Equation 41
  2.5.1 Inter-arrival times for reaction events 42
2.5.2 Index of reaction to be simulated
2.5.3 Inter-arrival times for duplication events
2.5.4 Inter-arrival time for division events
2.5.5 Volume ratio of the mother to the daughter cell
2.5.6 Number of non-chromosomal molecules inherited by one daughter cell
2.5.7 Number of chromosomal molecules inherited by one daughter cell
2.5.8 Pseudo-code

2.6 A Simple Model-Problem
2.6.1 Reaction Network, Growth and Division Mechanism
2.6.2 Numerical Solution of the M-Equation and MC Simulations

2.7 Conclusion

Chapter 3 : Extending the Modeling Approach to the Cell Population Level

3.1 Definitions
3.2 The Cell Population M-Equation
3.2.1 Probability Flow due to reaction
3.2.2 Probability Flow due to DNA duplication
3.2.3 Probability flow due to growth
3.2.4 Inflow due to division
3.2.5 Overall M-Equation
3.3 MC Simulation of the Cell Population M-Equation
3.3.1 Pseudo-Code
3.4 A Simple Model-Problem
3.4.1 Reaction Network; Growth, Duplication and Division Mechanisms
3.4.2 Simulation Results
3.5 Conclusion

Part II: Modeling Cell Population Heterogeneity in a Genetic Network with Positive Feedback Architecture

Chapter 4 : The lac operon: Review of Experimental and Modeling Studies
4.1 Experimental Studies
4.2 Modeling Studies
4.3 Conclusion

Chapter 5 : Quantifying the Effect of Intrinsic Noise for the lac operon Network at
List of Figures

Figure 1.1: Heterogeneous *E. coli* populations bearing the pTAK117 plasmid (Gardner et al. 2000). Picture taken with a fluorescence microscope (courtesy of Konstantinos Spectieris). ................................................................. 3

Figure 1.2: Panel (a): schematic representation of the interactions existing in the toggle genetic switch (pTAK117). Panel (b): illustration of the pTAK117 plasmid construct ................................................................. 9

Figure 1.3: Panel (a): schematic representation of the interactions existing in the repressilator genetic network, which is capable of exhibiting oscillatory behavior. Panels (b, c): illustration of the repressilator (b) and the corresponding reporter (c) plasmid constructs .................................................... 10

Figure 2.1: Explanation of the terms lineage and line as used in this work. Lineage is the set of all the daughter cells of a mother cell. Line is a collection daughter cells defined recursively: start from one mother cell; choose one of its daughter cells; set the daughter to mother cell; repeat. ......................... 25

Figure 2.2: Length to width ratio for a rod shaped bacterium. ............................................. 34

Figure 2.3: Panel (a): Transient behavior of the mean molecular content and the mean volume for the first 25 min (both computed by the transient solution of the M-Equation). Panel (b): Stationary probability distribution (t = 50 min) computed by solving the M-Equation. Panel (c): Representative stochastic paths for the molecular content and the volume for the first 25 min (both computed by MC simulation). Panel (d): Stationary probability distribution (t = 50 min) computed by multiple MC simulations and sampling at the end of the simulation ................................................................. 55

Figure 2.4: Convergence in the mean square sense. Panel (a): multiple MC runs with sampling in the final time (t = 50 min). Panel (b): single long MC simulation with sampling over time. In both panels the open circles denote the error calculated from simulations (using equation 2.71) and the line shows the theoretically expected behavior (equation 2.72) ................................................................. 57

Figure 3.1: Panel (a): Transient behavior of the Cfp content and the volume for a single cell line out of the cell population. Panel (b): Transient behavior of the
average Cfp content of the cell population. Panel (c): Number of cells in the population as a function of time. Panel (d): Normalized Yfp content with respect to the normalized Cfp content. Each point represents one cell. Color coding corresponds to density of points. Nominal parameter set (Table 3.1)........... 77

Figure 3.2: Panels (a - d) as in Figure 3.1 for the following parameter set: $k_1 = 4800 \text{ nM/min}, k_2 = 8500 \text{ nM/min}, k_5 = 0.12 (\text{nM-min})^{-1}, n_s = n_d = 10000$. The requirement $q \to \infty$ was numerically implemented by setting the daughter volumes equal to half the mother volume at every division. All other parameters as in Table 3.1............................................ 79

Figure 3.3: As in Figure 3.1d for the following parameter sets: panel (a), intrinsic noise only: $k_1 = 4800 \text{ nM/min}, k_2 = 8500 \text{ nM/min}, k_5 = 2\cdot 10^{-4} (\text{nM-min})^{-1}, n_s = n_d = 10000$. The requirement $q \to \infty$ was numerically implemented by setting the daughter volumes equal to half the mother volume at every division; panel (b), extrinsic noise arising only from fluctuations in the RNA polymerase: $k_1 = 1.2 \text{ nM/min}, k_2 = 8500 \text{ nM/min}, k_5 = 480 \text{ (nM-min)}^{-1}, n_s = n_d = 10000$. The requirement $q \to \infty$ was numerically implemented by setting the daughter volumes equal to half the mother volume at every division. All other parameters as in Table 3.1............................................ 80

Figure 3.4: As in Figure 3.1d for the following parameter sets: panel (a), $k_1 = 4800 \text{ nM/min}, k_2 = 8500 \text{ nM/min}, k_3 = 2000 \text{ nM/min}, k_4 = 2.4\cdot 10^{-5}$ (nM-min)$^{-1}, k_5 = 0.12 (\text{nM-min})^{-1}, n_s = n_d = 10000, q \to \infty$; panel (b): as in panel (a) with $k_5 = 0.12\cdot 10^{-3} (\text{nM-min})^{-1}, O_{cfp,Total} = O_{cftp,Total} = 1000$; panel (c): as in panel (a) with $k_3 = 8.1\cdot 10^{-2} \text{nM/min}, k_4 = 7.2\cdot 10^{-1} (\text{nM-min})^{-1}$; panel (d): as in panel (b) with $k_3 = 35 \text{nM/min}, k_5 = 0.12\cdot 10^{-3} (\text{nM-min})^{-1}, O_{cfp,Total} = O_{cftp,Total} = 1000$. Parameters that are not mentioned have values as in Table 3.1............................................ 82

Figure 5.1: Panel (a): schematic representation of the interactions taken into account in the $lac$ operon network. Panel (b): Interaction diagram for the key species of the $lac$ operon genetic switch (for species notation see Table 5.1; + and − denote positive and negative feedback respectively).................................... 97
Figure 5.2: Bifurcation diagrams with respect to the extracellular IPTG concentration $[I_{ex}]$ for the following variables: panel (a) the total LacY concentration; panel (b) the intracellular IPTG concentration; panel (c) the total repressor concentration (the dotted-dashed line corresponds to the limiting $[R]_T$ as calculated from equation AIV.46). Solid lines represent stable steady states and dashed lines unstable ones. Parameters as in Table 5.2. 105

Figure 5.3: Two parameter bifurcation diagrams for the turning points showing the effect of several biomolecular parameters on the bistable regime (enclosed between the curves). Shown are the effects of repressor-inducer complex degradation rate (a), of $lacI$ transcription rate (b), of repressor-inducer dissociation rate (c) and of total operator concentration (d). Parameters as in Table 5.2 except as noted. 108

Figure 5.4: Effect of the alternative derepression mechanisms transiently and asymptotically. Panels (a) and (b) correspond to the mechanism where inducer binds only to repressor-operator complex (thus, for the other mechanism, $k_{dr1}$ and $k_{dr2}$ are both zero): loss of bistability and faster induction in the case of slower repressor-operator association and dissociation rates. $[k_s \times k_r] \div 50$ denotes that both $k_r$, $k_r$ have values 50-fold lower than those of Table 5.2. Panels (c) and (d) correspond to the mechanism where inducer binds only to free repressor (thus, $k_{dr2}$ and $k_{dr2}$ are both zero): Shift of the bistable regime and slower induction for the case of slower repressor-inducer association and dissociation rates. The time course for $[k_{dr1} \times k_{dr1}] \div 500$ finally approaches the lower steady state. Panels (e) and (f) correspond to the mechanism where inducer binds to repressor-operator complex (thus, $k_{dr2}$ and $k_{dr2}$ are both zero): Broadening of the bistable regime and slower induction for the case of slower repressor operator complex-inducer association and dissociation rates. Parameters as in Table 5.2 except as noted. For all induction transients the IC is the corresponding steady state for $[I_{ex}] = 0 \, \mu M$ and at $t = 0$ min an extracellular IPTG concentration $[I_{ex}] = 100 \, \mu M$ is imposed. 111

Figure 5.5: Comparison of the stochastic average with the deterministic steady states. The error bars denote two standard deviations (i.e. the min and max of
the error-bars correspond to points $\mu \pm \sigma$). Parameters as in Table 5.2.................... 115

Figure 5.6: Typical sample path for the stochastically simulated $lac$ operon system for $[I_{ex}] = 20 \ \mu M$ (panel a), and Probability Mass Function for this parameter set (panel b). For panels (c) and (d) $[I_{ex}] = 100 \ \mu M$. Parameters as in Table 5.2.................................................................................................. 116

Figure 5.7: The stationary potential for $[Y]_T$ compared to the deterministic bifurcation diagram for a range of IPTG concentrations: effect of stochasticity on the behavior of the system in different parameter regimes. Panel (a) nominal parameter set (Table 5.2). Panel (b): 100-fold fast $lacY$ transcription ($k_{s1MY} = 50 \ \text{min}^{-1}$, $k_{s1MY} = 1 \ \text{min}^{-1}$) and slow translation ($k_{sY} = 0.3 \ \text{min}^{-1}$). Panel (c): 100-fold fast $lacI$ transcription ($k_{sMR} = 23 \ \text{min}^{-1}$) and slow translation ($k_{sR} = 0.15 \ \text{min}^{-1}$). Panel (d): simultaneous 100-fold fast $lacY$ and $lacI$ transcription and slow translation for both proteins ($k_{s1MY} = 50 \ \text{min}^{-1}$, $k_{s1MY} = 1 \ \text{min}^{-1}$, $k_{sMR} = 23 \ \text{min}^{-1}$, $k_{sY} = 0.3 \ \text{min}^{-1}$, $k_{sR} = 0.15 \ \text{min}^{-1}$). Panel (e): 10-fold slower repressor-operator association and dissociation with only the 1st derepression mechanism, considering fast transcription and slow translation for $lacY$ ($k_r = 96 \ \text{nM}^{-1} \cdot \text{min}^{-1}$, $k_{r-r} = 0.24 \ \text{min}^{-1}$, $k_{dr2} = 0 \ \text{nM}^{-2} \cdot \text{min}^{-1}$, $k_{dr2} = 0 \ \text{nM}^{-1} \cdot \text{min}^{-1}$, $k_{s1MY} = 50 \ \text{min}^{-1}$, $k_{s1MY} = 1 \ \text{min}^{-1}$, $k_{sY} = 0.3 \ \text{min}^{-1}$). Panel (f): 10-fold faster repressor-operator association and dissociation with only the 1st derepression mechanism, considering fast $lacY$ transcription and slow translation ($k_r = 9600 \ \text{nM}^{-1} \cdot \text{min}^{-1}$, $k_{r-r} = 24 \ \text{min}^{-1}$, $k_{dr2} = 0 \ \text{nM}^{-2} \cdot \text{min}^{-1}$, $k_{dr2} = 0 \ \text{nM}^{-1} \cdot \text{min}^{-1}$). Marked with $\bullet$--$\bullet$ is the stochastic bistable (bimodal) regime and with $\star$--$\star$ the deterministic. Surface color corresponds to stationary potential values and the bimodal regime is shaded green.

Parameters as in Table 5.2 unless otherwise noted.............................................................. 118

Figure 5.8: Panel (a): Mean First Passage Times (MFPTs) for a transition of $[Y]_T$ from the lower to the upper attracting vicinity (circles) or opposite (triangles) for a range of extracellular IPTG concentrations. Fast $lacY$ transcription and slow translation were considered. The min and max of the error-bars correspond to the MFPT $\pm$ one standard deviation. Superimposed for comparison is the corresponding deterministic bifurcation diagram for $[Y]_T$,
with arbitrarily scaled y-axis. Panel (b): PDF for the First Passage Time $T_{\text{low-sup}}$ for $[I_{\text{ex}}] = 20 \ \mu\text{M}$. The Mean First Passage Time is $1/\lambda \approx 179 \ \text{min}$. Parameters as in Figure 5.7b. ................................................................. 123

Figure 5.9: Panel (a): Autocovariance for extracellular IPTG $[I_{\text{ex}}] = 10 \ \mu\text{M}$. The autocovariance function decays exponentially fast, with an autocorrelation time $\tau_e \approx 25 \ \text{min}$ for this case. Panel (b): dependence of the autocorrelation times on the induction levels as quantified by $[I_{\text{ex}}]$. Superimposed for comparison is the corresponding deterministic bifurcation diagram for $[Y]^T$, with arbitrarily scaled y-axis. Parameters as in Figure 5.7b. ................................................................. 125

Figure 6.1: Panel (a): Transients for all the cells in the population starting from a single cell close to the lower state. Panel (b): Dynamics of the population mean and standard deviation for the total LacY concentration. Panels (c, d): As in panels (a, b) starting with a cell close to the upper state. $[I_{\text{ex}}] = 22 \ \mu\text{M}$ and all other parameters as in Table 5.2 and Table 6.1................................................................. 134

Figure 6.2: Panel (a): Dynamics of the population average for the total LacY concentration. The simulation was initialized with a single cell close to the lower attracting state. The black curve shows the number of cells in the population and the vertical dashed red line shows the time when the Constant Number MC Scheme was triggered. The maximum allowed number of cells in the population was $10^4$. Panel (b): As in panel (a) with the initial cell close to the upper attracting state. Panels (c, d): Timecourses for single cells: only one daughter is followed after division, and thus a line of cells is simulated, see also Figure 2.1. The cell was initialized close to the lower (c) or upper (d) attracting state. No transitions between attracting states are observed even for large times. $[I_{\text{ex}}] = 22 \ \mu\text{M}$ and all other parameters as in Table 5.2 and Table 6.1................................................................. 135

Figure 6.3: Panel (a): NDF for the total LacY concentration for $[I_{\text{ex}}] = 10 \ \mu\text{M}$ at time $t = 446.3 \ \text{min}$ where the number of cells was $N = 47826$. Panel (b): As in panel (a) but for the logarithm of the total LacY concentration (this is what a distribution obtained by FC with logarithmic amplification would look like). Panel (c): Scatter plot for the volume of the cells versus the total LacY content
xiii

corresponding to panel (a) (this is what a FSC-FL1 scatter plot obtained by FC with logarithmic amplification in both channels would look like). Panels (d, e, f): As in panels (a, b, c) with \([I_{ex}] = 20 \ \mu M\), at time \(t = 446.1 \ \text{min}\) where the number of cells was \(N = 47595\). Panels (g, h, i): As in panels (a, b, c) with \([I_{ex}] = 50 \ \mu M\), at time \(t = 446.1 \ \text{min}\) where the number of cells was \(N = 47611\). All other parameters as in Table 5.2 and Table 6.1.

Figure 6.4: Panels (a, b): Dynamics of the population mean for the membrane area over the volume ratio and the total operator concentration. Panel (c): agreement between the cell population model and the single cell model when the average values for \(A/V\) and \([O_T]\) are used. For the blue curve, the average values for \(A/V\) and \([O_T]\) were taken to be as shown in panels (a) and (b). For the red curve, the average values were taken to be equal to the estimates of equations (6.30, 31). Panel (d): a representative bimodal NDF and the corresponding averages and standard deviations for \([I_{ex}] = 20 \ \mu M\). Parameters as in Table 5.2 and Table 6.1.

Figure 6.5: Panels (a, b): Transient dynamics of the population mean for the switching from the low \((I_{ex} = 0 \ \mu M)\) to the high state \((I_{ex} = 60 \ \mu M)\) (panel a) and conversely (panel b). Nominal parameter set. For the blue curve (single cell 1), the average values for \([O_T]\) and \(A/V\) were taken to be as shown in Figure 6.4a, b, and for the red curve (single cell 2), they were calculated from equations (6.30, 31). Panel (c): As in panel (a) but with \(k_{s0MY} = 0.25 \cdot 10^{-1} \ \text{min}^{-1}\), \(k_{s1MY} = 0.005 \cdot 10^{-1} \ \text{min}^{-1}\), \(\lambda_{MY} = 0.001155\). Panel (d): The distribution of the division times (for comparison with the switching times). Parameters as in Table 5.2 and Table 6.1 unless otherwise noted.

Figure 6.6: Panel (a): timecourses of the cells in a population. The simulation accounts for stochastic reaction occurrence. Panel (b): the population average from the simulation of panel (a). \([I_{ex}] = 20 \ \mu M\) and nominal parameter set (Table 5.2, Table 6.1).

Figure 6.7: Panel (a): the population average calculated from the simulation batch. Panel (b): Number of cells in each of the 20 batches versus time. \([I_{ex}] = 20 \ \mu M\) and nominal parameter set (Table 5.2, Table 6.1).
Figure 6.8: Comparison of the NDFs for deterministic ("Deter. Rxn") versus stochastic ("Stoch. Rxn") reaction dynamics. For all deterministic simulations the cell population was initiated with 20 cells and $N_{\text{cells max}} = 10000$. For all stochastic simulations batches of 20 simulations were run. In all cases sampling was performed at $t = 300$ min. Panel (a): $[I_{\text{ex}}] = 10 \, \mu\text{M}$. Panel (b): $[I_{\text{ex}}] = 20 \, \mu\text{M}$. Panel (c): $[I_{\text{ex}}] = 30 \, \mu\text{M}$. Panel (d): $[I_{\text{ex}}] = 50 \, \mu\text{M}$. Nominal parameter set (Table 5.2, Table 6.1).

Figure 6.9: Panel (a): Comparison of the single cell probability distribution function with the cell population NDF. For the population distribution, a batch of 10 simulations was run with $N_{\text{cells max}} = 500$, $[I_{\text{ex}}] = 30 \, \mu\text{M}$ and nominal parameter set (Table 5.2, Table 6.1), and sampled at $t = 270$ min. For the single cell simulation a cell line was tracked ($N_{\text{cells max}} = 1$) for $10^5$ min and samples were taken periodically in time with $\Delta t = 10$ min. Panel (b): as in panel (a) with $[I_{\text{ex}}] = 30 \, \mu\text{M}$ and 100-fold faster $lacY$ transcription ($k_{s1MY} = 50 \, \text{min}^{-1}$, $k_{s1MY} = 1 \, \text{min}^{-1}$) and slower translation ($k_{sY} = 0.3 \, \text{min}^{-1}$) and $[I_{\text{ex}}] = 20 \, \mu\text{M}$. The simulation batch consisted of 20 simulations and was sampled at $t = 250$ min. The cell line was tracked for $10^5$ min of simulated time.

Figure 7.1: Schematic for the cell movement model.
List of Tables

Table 1.1: Summary of the sources of heterogeneity taken into account by the
major theoretical frameworks ................................................................. 20
Table 2.1: Parameters for the model problem and the simulators .................. 54
Table 3.1: Parameter values for the two promoter system (no repressive action) .... 76
Table 5.1: lac operon species notation ..................................................... 96
Table 5.2: Parameters of the lac operon reaction model ............................. 101
Table 5.3: Propensity functions for the stochastic lac operon model ............ 114
Table 6.1: Population parameters for the lac operon ................................ 133
Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>BC</td>
<td>Boundary Condition</td>
</tr>
<tr>
<td>CFP</td>
<td>Cyan Fluorescent Protein</td>
</tr>
<tr>
<td>CPB</td>
<td>Cell Population Balance</td>
</tr>
<tr>
<td>DNA</td>
<td>Deoxyribonucleic Acid</td>
</tr>
<tr>
<td>FC</td>
<td>Flow Cytometry</td>
</tr>
<tr>
<td>GFP</td>
<td>Green Fluorescent Protein</td>
</tr>
<tr>
<td>IC</td>
<td>Initial Condition</td>
</tr>
<tr>
<td>M-Equation</td>
<td>Master Equation</td>
</tr>
<tr>
<td>MC</td>
<td>Monte Carlo</td>
</tr>
<tr>
<td>mRNA</td>
<td>messenger Ribonucleic Acid</td>
</tr>
<tr>
<td>NDF</td>
<td>Number Density Function</td>
</tr>
<tr>
<td>ODE</td>
<td>Ordinary Differential Equation</td>
</tr>
<tr>
<td>PCR</td>
<td>Polymerase Chain Reaction</td>
</tr>
<tr>
<td>PDF</td>
<td>Probability Density Function</td>
</tr>
<tr>
<td>RFP</td>
<td>Red Fluorescent Protein</td>
</tr>
<tr>
<td>RNA</td>
<td>Ribonucleic Acid</td>
</tr>
<tr>
<td>RT-PCR</td>
<td>Reverse Transcription Polymerase Chain Reaction</td>
</tr>
<tr>
<td>SDE</td>
<td>Stochastic Differential Equation</td>
</tr>
<tr>
<td>SVNMC</td>
<td>Stochastic Variable Number Monte Carlo</td>
</tr>
<tr>
<td>YFP</td>
<td>Yellow Fluorescent Protein</td>
</tr>
</tbody>
</table>
Part I: Developing Mathematical and Computational Tools for Integrating the Major Sources of Cell Population Heterogeneity
Chapter 1:

Introduction

This thesis focuses on the link between stochasticity and heterogeneity. Therefore, in this introductory chapter we first give some definitions and subsequently review the pertinent literature. We further present the objectives of our study and give an outline of the thesis.

1.1 The Study of Heterogeneity: Motivation

Any living organism can be characterized by its *genotype*, namely the DNA sequence encoding the organism's hereditary information. The genotype, along with environmental stimuli, shapes the observable properties of the organism which are called the *phenotype*. A population in which all individuals have the same genetic content is referred to as an *isogenic* population. The term heterogeneity, as used in the current work, pertains to the phenotypic differences between isogenic cells which live in the same environment. Note the duality of this restriction: uniformity applies to both the genetic content and environment; however, other characteristics, for example the protein content, can vary between individuals. This potential difference will in fact contribute to the shaping of
heterogeneity; we will elaborate more on this point in a subsequent section.

Heterogeneity is observed in virtually every population of living organisms; however, this work focuses on heterogeneous *Escherichia coli* populations (Figure 1.1). On the one hand, this choice illustrates that even the simplest forms of life, bacteria, can exhibit surprisingly complex behavior. On the other hand, trying to study heterogeneity between individual multicellular organisms would be an extremely complicated task, since a multicellular organism itself is composed of interacting heterogeneous populations of cells. Thus, numerous interactions, leading to the development of the organism, would have to be taken into account. This complexity could easily lead to an intractable problem if one did not try to study the simpler case of one heterogeneous cell population first.

Furthermore, the existing profound knowledge about *E. coli* makes this organism a perfect prototype system, in which the effects of several factors on the observed heterogeneity are to be investigated.

The study of cell population heterogeneity is of paramount importance for two main reasons. First, it will serve as the basis for rigorously designing biochemical processes in which bacterial populations produce a desired product (for example nutritional substances or drugs). Until the previous decade, the biological paradigms and the state-of-the-art modeling frameworks consistently neglected heterogeneity (Chung and Stephanopoulos.
1995; Fedoroff and Fontana 2002). The fundamental assumption of such theoretical investigations was that all cells behave like the average cell, and thus their behavior could be described by continuum models consisting of ordinary differential equations (ODEs). However, this behavior is not what is observed in vivo, since cell populations are inherently heterogeneous (Avery 2006; Davidson and Surette 2008). More importantly, even if one is interested only in the average dynamics, it turns out that use of continuum models will result in incorrect predictions (McAdams and Arkin 1998). Therefore, one has to explicitly account for the heterogeneous nature of the cell population in order to correctly predict the production and the productivity as well as to effectively design and/or control a biochemical process (Mantzaris 2005b).

The second reason for studying non-genetic heterogeneity is that it may have physiological importance in the survival of cell populations. It has been proposed that the viability of the cell population and the efficient adaptation to sudden changes in environmental conditions may be linked to the phenotypic heterogeneity (McAdams and Arkin 1999; Sumner and Avery 2002; McAdams et al. 2004; Veening et al. 2008a; Veening et al. 2008b). Thus, phenomena such as the resistance of certain infectious bacteria to antibiotics could be explained on the basis of the existence of a small subpopulation that overcomes the shock and continues growing afterwards (Booth 2002). Moreover, even if the environmental changes do not pose a threat for the viability of the cell population, it has been demonstrated theoretically that a heterogeneous cell population can achieve faster growth rates than those of a homogeneous one (Thattai and van Oudenaarden 2004).

We therefore see that the inherent heterogeneity of a cell population can result in
phenomena that homogeneous models are unable to capture. A systematic exploration of these phenomena is indispensable for understanding the dynamics of cell populations and developing practical applications. Thus, in the subsequent sections, we will review the experimental and theoretical studies aiming towards this goal.

1.2 Historical Perspective

In science one cannot isolate experiments from theory, since they interact in a non-linear fashion; sometimes they progress together, other times one of the two gets ahead of the other, and frequently they get trapped in recurring loops! Even so, for methodological purposes we chose to break this review into two different sections, one for experimental approaches and one for theoretical investigations.

1.2.1 Experimental Approaches

The presence of heterogeneity is not a newly discovered phenomenon. Non-hereditary phenotypic variations in prokaryotic cell populations have been experimentally documented and characterized for over half of a century. Delbrück (1945) was one of the first to characterize a heterogeneous response in an isogenic cell population. His experiments involved infection of *E. coli* with a bacteriophage and measurement of cell lengths and number of released viral particles upon lysis (burst) of each cell (burst size distribution). The results of this study indicate that phenomena related to the intracellular viral growth must play a significant role in shaping the observed variability (similar experiments were later performed by Kourilsky 1973). It is interesting to note that even though Delbrück did not know the specifics of the viral growth process, he postulated that the mechanism should be autocatalytic based on his noticing large fluctuations in the
number of released viral particles, a key observation for the current work as well.

Stocker (1949), and later Lederberg and Iino (1956), described variability in bacterial flagellar phases in *Salmonella typhimurium* populations. In the latter study, although the focus is on the genotypic source of this variability, the cytoplasmic composition is also mentioned as a possible source. The argument, cited from Ephrussi (1953), is that “unless development involves a rather unlikely process of orderly and directed gene mutation, the differential must have its seat in the cytoplasm.” This comment hints towards viewing heterogeneity as a partial result of the random nature of biochemical processes occurring in the cytosol.

The link of randomness with heterogeneity was later explicitly put forward by Powell (1958) in the context of his work about variation in bacterial generation times (Tyson and Hannagen 1985; Bremer 1986). He attributes this variation to two factors: chance, as he mentions “molecular accidents,” and heredity. Furthermore, the intracellular contents are partitioned unevenly amongst the two daughter cells upon division of the mother cell (Tyson and Hannagen 1985). This phenomenon, termed *unequal partitioning*, further contributes to heterogeneity (Eakman et al. 1966; Tsuchiya et al. 1966; Fredrickson et al. 1967) and will be accounted for in the current work. In the following we will also use the term *stochastic partitioning* to refer to non-predictable partitioning that may or may not be even.

Heterogeneous responses in the cell population level have also been observed in a naturally occurring genetic network, which will serve as a prototype for our work, the *lac* operon genetic switch. Benzer (1953) and subsequently Novick and Weiner (1957) and Maloney and Rotman (1973) reported variability in the β-galactosidase intracellular
concentrations. This variability has an “all-or-none” character (Novick and Weiner 1957): two distinct subpopulations of cells coexist, one with high yield of β-galactosidase and the other with low. Similar results have been reported by using in situ Reverse Transcription Polymerase Chain Reaction (RT-PCR) to monitor lac messenger Ribonucleic Acid (mRNA) (Tolker-Nielsen et al. 1998).

Non-genetic variability has also been observed in the tumbling frequencies of bacteria (Spudich and Koshland Jr 1976; Koshland Jr 1978). In absence of chemotactic stimuli, a bacterium shows brief periods of smooth swimming interrupted by rapid irregular motion (tumbling) which reorients its swimming direction. By analyzing tumbling frequencies in isogenic cell populations, Spudich and Koshland demonstrated that there is a large individual variability not attributable to genetics or the “age” of the bacterium (position in cell cycle); it rather seems to be due to random events in the bacterium’s life. The researchers proposed that the probabilistic fluctuations in the small numbers of regulatory molecules at the single cell level lead to the observed Poisson distribution of tumbling frequencies at the population level. They also argued that non-genetic heterogeneity aids in the survival of a population exposed to widely varying conditions during its lifetime.

Furthermore, Chung and Stephanopoulos (1995) have shown cell-to-cell variability in sporulating cultures of Bacillus subtilis. By using fusions of sporulating genes with lacZ, whose product could be detected with flow cytometry (FC), it was demonstrated that within such cultures there exist two distinct subpopulations: an induced population, which has activated the developmental pathway for spore formation and an uninduced population, in which the aforementioned pathway is inactive. Chung and Stephanopoulos (1995) stressed the inadequacy of deterministic single cell models to explain the cell
culture behavior and underlined the importance of taking into account the heterogeneous nature of the cell culture, an idea which we intend to further support with the current work. More recent works have established the *Bacillus subtilis* species as a paradigm for non-genetic differentiation in bacteria (Dubnau and Losick 2006; Iber et al. 2006; Aguilar et al. 2007; Lopez et al. 2008).

In the aforementioned works, naturally occurring networks were primarily studied with few or no modifications. In contrast to this approach, recent studies aim towards analyzing heterogeneity within bacterial cultures which incorporate artificial genetic networks (Kærn et al. 2003; for reviews see Kærn et al. 2005; Raser and O'Shea 2005; Sprinzak and Elowitz 2005; Smits et al. 2006). This revolutionary approach has several advantages over characterizing naturally occurring genetic networks. First, the ability to design a network at will allows for a methodical analysis of the effect of various biomolecular mechanisms on the observed cell population heterogeneity. Furthermore, analysis of the cells' phenotype is greatly facilitated, since reporter proteins can be integrated in these networks. Therefore, the expression of a gene of interest can be conveniently and reliably measured with high-throughput techniques, such as FC. Finally, the researcher is given the opportunity to rigorously link the experimental observations with mathematical models, since precise genetic architectures can be experimentally probed and theoretically analyzed at the same time.

Thus, Gardner et al. (2000) constructed an artificial genetic toggle switch with two genes, *lacI* and *cltS*, under the influence of inducible promoters such that one is repressing the production of the other (Figure 1.2). By incorporating the GFP-mut3 variant of the Green Fluorescent Protein (GFP) as a reporter co-transcribed with *clt*,
Figure 1.2: Panel (a): schematic representation of the interactions existing in the toggle genetic switch (pTAK117). Panel (b): illustration of the pTAK117 plasmid construct.

Gardner et al. (2000) demonstrated heterogeneity in a bistable genetic network. Interestingly, they observed that at medium induction levels the cell population split into an induced and an uninduced subpopulation. This result is similar to that for the lac operon switch and seems to generalize for bistable networks.

Furthermore, Besckei and Serrano (2000) engineered a genetic network with negative feedback architecture in which a transcription factor, a fusion of the tetracycline repressor (TetR) with Enhanced GFP (E-GFP), negatively regulates its own production. By varying the strength of the negative feedback loop, Besckei and Serrano showed that stronger negative feedback decreases the observed phenotypic variability (heterogeneity) of the cell population.

Later, Besckei and coworkers (2001) also studied the case of a genetic network with positive feedback architecture by constructing synthetic gene circuits involving the tetracycline-responsive trans activator (rtTA) in the eukaryotic Saccharomyces cerevisiae. The main finding of this study is that the positive feedback mechanism is responsible for the all-or-none phenotypic response of the system. In particular, if the constitutive expression of a transcriptional activator follows a continuous gradient, the
positive feedback results in a phenotypic bistable behavior, thereby transforming a graded signal into a binary switch. The noise-induced transitions between the states of the switch are responsible for the more pronounced heterogeneous phenotypic responses; thus, positive feedback increases the observed phenotypic variability.

Bistability has also been observed in prokaryotic cells for the case of another network with positive feedback: in a study by Isaacs et al. (2003) the bacteriophage $\lambda$ genetic switch was isolated and incorporated into *E. coli* cells. However, in this network, bistability is observed only after introducing a mutation of the $cl$ gene that renders the CI repressor temperature-sensitive.

A network with more complex architecture, termed the repressilator, was also constructed by Elowitz and Leibler (2000). This network consists of three genes; $tetR$, 


Figure 1.3: Panel (a): schematic representation of the interactions existing in the repressilator genetic network, which is capable of exhibiting oscillatory behavior. Panels (b, c): illustration of the repressilator (b) and the corresponding reporter (c) plasmid constructs.

lacI and cl, that cyclically repress the production of each other (Figure 1.3).
Consequently the network exhibits oscillations even though none of the aforementioned
genes is part of any natural biological clock. However, the oscillations do not remain
synchronized, possibly because of the inherent noise. A different study in a naturally
occurring circadian oscillator in *Synechococcus elongatus* implies that synchrony can be
ensured not only by intercellular interactions, but also by the dynamics of the intracellular
network (Mihalcescu et al. 2004). Thus, the occurrence of oscillatory networks hints at
the need for an amendment of the notion of heterogeneity, to include cases where the cell
population characteristics are heterogeneous over time in addition to being heterogeneous
over the population ensemble.

Apart from the oscillatory dynamics, excitability can also induce heterogeneity, as
shown by Süel et al. (2006). In this study, key genes of the competence pathway were
monitored with fluorescent markers in an engineered *Bacillus subtilis* strain. It was
shown that the excitable properties of the genetic network can, in presence of
stochasticity, result in a random activation of the competence pathway, or trigger the
reverse process, namely the switching back to the vegetative growth pattern. Such
phenomena result in temporal heterogeneity at the cell population level.

The aforementioned studies aimed primarily to elucidate the effect of genetic
architecture in the observed phenotypic heterogeneity. The need for a greater depth of
understanding, though, inspired studies which aimed at assessing the effect of
stochasticity at each step of the protein expression mechanism (transcription versus
translation). To this end, Ozbudak et al. (2002) performed studies in which the rates of
transcription and translation for a fluorescent reporter gene in the chromosome of *B.*
*subtilis* were independently varied. Subsequently, the observed phenotypic heterogeneity was quantified, and it was found that the main source of the increased heterogeneity was the increase of translational efficiency. On the contrary, similar studies in a eukaryotic system in *S. cerevisiae* showed that both factors contribute to the observed heterogeneity (Blake et al. 2003).

Other studies attempted to distinguish between different sources of noise. To do so, Elowitz et al. (2002) pioneered the construction of a genetic network with two different reporters, Yellow Fluorescent Protein (YFP) and Cyan Fluorescent Protein (CFP), under the influence of identical promoters and measured the levels of fluorescence of each cell in a population sample. This enabled the distinction between *intrinsic noise*, which creates differences in the yellow and cyan fluorescence levels of the same cell, and *extrinsic noise*, which creates differences between the fluorescence levels of different cells. The experiment of Elowitz et al. showed that both sources of noise contribute to the overall heterogeneity, but extrinsic noise seems to be more pronounced. Moreover, Rosenfeld et al. (2005) showed that the autocorrelation time of extrinsic noise is 40 ± 10 min, a value that compares well with the duration of the cell cycle (45 ± 10 min), whereas intrinsic noise has a smaller autocorrelation time (< 10 min). This difference implies that effects such as the variability in cell division times or the stochastic partitioning events contribute to the extrinsic noise, whereas the stochasticity in the intracellular reactions contribute to the intrinsic noise (Rosenfeld et al. 2005).

In addition, recent studies have undertaken the intriguing task of analyzing the propagation of noise into more complex architectures, involving several regulatory cascades. Hooshangi et al. (2005) demonstrated how the length of transcriptional
regulatory cascades affects the propagation of noise in gene expression by constructing networks in which a YFP is finally produced after one, two, or three repression stages. They observed that longer cascades result in sharper switching behavior. Furthermore, they reported high population heterogeneity at intermediate transcriptional induction levels; in these conditions each additional regulatory step was found to roughly double the expression noise. Therefore, in this case, the more pronounced switching behavior comes at the expense of increased heterogeneity and loss of synchrony in a cell population.

All the aforementioned experimental works revealed that there are several biological factors contributing to the non-genetic heterogeneity observed in cell populations growing in homogeneous environments. Moreover, the pattern of variability expressed in the phenotype of the population is a result of the complex interplay between the several sources of noise and the specific genetic architecture. Understanding this interplay is an intriguing task which cannot be performed by experimental work alone; mathematical modeling can obviously help in elucidating the underlying dynamics and reveal the links between cause and effect in these systems. To this end, several theoretical works have appeared in the literature, the most important of which we review in the next section.

1.2.2 Theoretical Approaches

In this section we first give a brief review of the major contributions in modeling cell population heterogeneity. Subsequently, we will discuss recent developments pertaining specifically to the use of stochastic models to describe heterogeneity, which is the focus of our work.

Attempts to model cell population heterogeneity have appeared in the literature as
early as '30s, when Otto Rahn and later David G. Kendall suggested two different models for the variability of cell division times (Rahn 1932; Kendall 1948; Kendall 1952), which were subsequently compared to experimental data by Powell (1958). These attempts are based on describing the statistics of the processes required for and culminating in cell division. However, they do not constitute a general framework for modeling phenotypic heterogeneity in cell populations.

A general method for modeling heterogeneity, though, was introduced much later (in the '60s) by Fredrickson and coworkers who developed the Cell Population Balance (CPB) modeling approach (Eakman et al. 1966; Tsuchiya et al. 1966; Fredrickson et al. 1967). The CPB models require single cell information to predict the distribution of phenotypic characteristics in the cell population level. In particular, the single cell information is summarized in the so-called intrinsic physiological functions, which give the growth rate, the division rate and the partition probability density with respect to the physiological state of the cell. By assumption, the physiological state fully describes the cell's status and is generally a vector, whose components can include the intracellular contents of chemical species as well as the morphometric characteristics of the cell. Therefore, in CPB models, heterogeneity is a consequence of the physiological functions, and its extend depends on their functional form, which account for the different growth and division rates of the cells, as well as for unequal partitioning effects.

Multidimensional CPB models are very difficult to solve, even with the current computational power available. Therefore, Monte Carlo (MC) algorithms were developed to simulate realizations of the underlying processes, and thus compute phenotypic distributions, numbers of cells or any other desirable characteristic of the cell population.
Shah et al. developed such an algorithm to simulate mass distribution dynamics (Shah et al. 1976) which was later extended by Hatzis et al. (1995) to simulate the multi-staged growth of phagotrophic protozoa. These algorithms are computationally intensive because the number of cells in the population increases exponentially in time; thus were developed constant-number MC algorithms that simulate a constant number of cells which is a representative sample of the overall population. As a matter of fact, these algorithms do not follow the dynamics of population growth. Mantzaris (2006) has developed a variable number Monte-Carlo algorithm in which the sample size can be increased up to a maximum. Hence, a single cell can be simulated up to the point that there are $N_{\text{max}}$ offsprings, from that point on the algorithm behaves as a constant-number (Smith and Matsoukas 1998).

However, CPB models are deterministic, in the sense that they neglect the discrete nature of (i) the cell population (Ramkrishna 2000) and (ii) the intracellular content. Therefore, these models do not account for stochastic effects originating from the low numbers of cells in the population; such effects are significant during the initial times of population growth and can be successfully simulated by a Variable Number MC algorithm proposed by Mantzaris (2006). Furthermore, CPBs cannot capture the inherent stochasticity of chemical reactions occurring in cellular control-volumes and, to the author's knowledge, no approach has been proposed for the coupled description of the following sources of heterogeneity: stochastic DNA duplication and division and stochasticity of intracellular reactions.

A conceptually different approach involves the use of the so-called ensemble models, introduced by Shuler and coworkers (Domach and Shuler 1984; Ataai and Shuler 1985;
see also Henson 2003 for a review on CPBs and ensemble models). These models are based on randomly perturbing the intracellular parameters or the Initial Conditions (ICs) of a single cell model and simulating each of the perturbed single cell models individually, thereby creating an ensemble. Thus, one can obtain distributions over the ensemble for any variable of the single cell model. Cell division can be accounted for in these models; however, stochastic partitioning was not incorporated in the original work by Domach and Schuler (1984). Heterogeneity is therefore realized via the single-cell model parameters (Henson 2003). The advantages of the ensemble models over the CPBs are mainly the capability of directly incorporating the single cell model into the ensemble model, the ability to solve problems involving many species and the simpler, in comparison to the CPBs, formulation which does not require knowledge of the physiological functions. The disadvantages of the ensemble models include the prohibitively slow dynamic simulation for large ensembles, coupled with the fact that the distribution’s resolution depends on the ensemble size. Furthermore, similarly to the CPBs, the ensemble models neglect the discrete nature of the intracellular content and do not take into account stochasticity of reaction phenomena.

Therefore, to account for the stochasticity in (i) the chemical reaction processes and (ii) the cell division events, stochastic models had to be developed. Not all models incorporate both stochastic factors and in some models, noise is imposed externally and thus artificially, as we will discuss in the following.

The idea that the occurrence of reactions is a stochastic process had already been introduced in the early 20th century (for review, see McQuarrie 1967), and early studies showed the effect of stochasticity in biochemical processes such as protein synthesis
(Singh 1969; Rigney and Schieve 1977). On the other hand, the effect of stochasticity in partitioning events had been demonstrated by Berg (1978), under the assumption of binomial partitioning (each protein molecule has equal probability of being inherited by either daughter cell). Furthermore, a model whose derivation is based on the random timings of bindings and dissociations of a transcription factor was presented by Ko (1991; 1992). Ko’s model describes stochasticity in gene induction rather than in protein synthesis.

An approach that takes into account stochastic effects in the entire biochemical pathway (gene induction and protein synthesis), but not in cell division events, was brought to attention later by McAdams and Arkin (1997). It is based on Gillespie’s MC algorithm (1976; 1977) which can be used to simulate exact sample paths of the Chemical Master Equation (M-Equation). The latter essentially describes the evolution of the probability of finding \( n_i \) molecules of species \( i = 1, \ldots, m \) in the control volume where the reactions into consideration occur. Thus, McAdams and Arkin (1997) used Gillespie’s algorithm to simulate the stochastic dynamics of intracellular processes, thereby showing that randomness can result in phenotypic variability within a cell population. An important consequence of taking into account stochasticity was the potential to explain experimental observations in which subpopulations with distinct phenotypic characteristics emerge from an isogenic initial population. Since the theory of stochastic dynamics predicts noise-induced transitions between attracting states, the observations just noted could be explained in terms of the stochastically driven “jumps” exhibited by a multistable genetic network. Arkin et al. (1998) further analyzed the lysis-lysogeny pathway of the \( \lambda \)-bacteriophage and compared the results obtained from the
stochastic model to experimental results for the fraction of infected cells selecting the lysogeny pathway. Additionally, Thattai and van Oudenaarden (2001) investigated theoretically the effect of network architecture in the cell population distributions and found that for a single gene, the noise is determined at the translational level and that negative feedback decreases the noise strength. Both findings compare well with the experiments performed by Ozbudak et al. (2002).

Gillespie’s algorithm was originally developed for a control volume that remains constant but Gardiner (1983) has showed how to modify the propensity functions to use the algorithm with a volume that changes with time. Gibson and Bruck (2000) came up with an exact efficient version of the algorithm that was extended by Swain et al. (2002) to account for cell growth and division. This latter algorithm accounts for linear single cell growth, division after fixed time T into two cells of equal sizes, binomial partitioning of the contents to the two daughters and DNA duplication at time arbitrarily set to 0.4·T. Only one daughter is followed after division. Using this algorithm Swain et al. (2002) demonstrated that the total noise of a genetic network can be decomposed into an intrinsic and an extrinsic component which have orthogonal contributions to the total noise. Extrinsic noise stems from noisy “inputs” to the genetic network such as a repressor concentration or the cell cycle state. Intrinsic noise stems from the randomness in the occurrence of the reactions of which the networks comprises. To test the orthogonality experimentally, they suggested the two reporter method, used in a subsequent experimental work (Elowitz et al. 2002).

Exponential cell growth and symmetric division was also incorporated to the Gillespie algorithm by Lu et al. (2004) who used the resulting algorithm in conjunction with hybrid
simulation techniques to analyze the behavior of an unregulated gene system. The two
aforementioned algorithms do not take into account variability in DNA duplication or
division times, and they assume that the mother cell produces two daughters with the
same volume. Furthermore, the hybrid simulation techniques used by the algorithm by Lu
et al. (2004) are valid only for the limiting case of small and fast noise which requires
high species copy numbers. Finally, both algorithms simulate single cells instead of cell
populations.

Gillespie’s algorithm is computationally intensive when the species into consideration
have high copy numbers. Thus, in some studies, stochasticity was modeled as an external
noise source by the Stochastic Differential Equations (SDEs) formulation (we will refer
to the use of SDEs as the Langevin approach following van Kampen 1992). In particular
Hasty and coworkers presented a model for the λ-bacteriophage genetic network (Hasty
et al. 2000; 2001), showing how random fluctuations can be used to control the state of a
biochemical switch. In this approach though, stochasticity does not stem from the
randomness in reaction occurrences or cell division events, thereby being somewhat
artificial. On the contrary, Kepler and Elston (2001) derived elegant approximations to
exact stochastic models describing gene-regulatory networks. In these approximate
models, noise is not imposed artificially but captures the inherent stochasticity of the
network in the limit of small noise amplitudes and fast fluctuations. Using these
approximations Kepler and Elston (2001) showed that qualitative changes in the
Probability Density Functions (PDFs) obtained by such networks can result solely from
changes in the rate of operator fluctuations.

None of the aforementioned algorithms simulates stochasticity in intracellular
reactions as well as in cell division at the cell population level. To this end, Mantzaris (2007) proposed a stochastic variable number Monte Carlo (SVNMC) algorithm that takes into account these sources of stochasticity. This algorithm is based on the deterministic analogue; the only difference is that instead of deterministic reaction expressions, SDEs are used (Langevin approach). Using this algorithm Mantzaris (2007) simulated a genetic network with positive feedback and showed that different sources of stochasticity can have a marked effect on the region of the parameter space where the system exhibits bistability. However, the Langevin approach neglects the discrete nature of the molecular content of the cells, as the copy numbers of species are treated as continuous variables. In fact the Langevin approach holds true for limiting cases of fast and small stochastic fluctuations since it is derived as an asymptotic approximation for large species copy numbers and fast operator fluctuations (Kepler and Elston 2001). Thus, the predictive power of the algorithm developed by Mantzaris (2007) is limited, since significant intrinsic noise is brought about by low species copy numbers which result in slow and large stochastic fluctuations. Finally, DNA duplication and the partitioning of molecules as discrete entities are not account for in this algorithm.

Table 1.1: Summary of the sources of heterogeneity taken into account by the major theoretical frameworks

<table>
<thead>
<tr>
<th>Source of Heterogeneity</th>
<th>Intrinsic Noise</th>
<th>Perturbed Intracellular Parameters</th>
<th>Growth</th>
<th>DNA Species</th>
<th>Stochastic Division</th>
<th>Population Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cell Population Balances</td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Chemical Master Equation</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Ensemble Methods</td>
<td></td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>SVNMC Algorithm</td>
<td>✓</td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>
1.3 Current Objectives

From the previous discussion, it emerges that none of the current mathematical frameworks accounts for all the various sources of heterogeneity, namely growth rate variability, stochasticity in the cell cycle events (such as DNA duplication for bacteria), and stochastic reaction occurrences. Furthermore, some of these frameworks simulate single cells instead of cell populations. Thus, the scope of part I of the current work is dual: (i) to develop a generic mathematical formulation that can incorporate the major sources of stochasticity at the single cell level and (ii) to extend the formulation to the population level.

The importance of developing such a formulation becomes apparent in view of the complex interplays of the different sources of heterogeneity observed in biological systems. On the one hand, studying in isolation one source of heterogeneity can definitely give insight on the effect of this source; for example, a Gillespie simulation can show the effect of stochasticity in intracellular reactions. On the other hand though, the sources of heterogeneity are in dynamical interaction. In our example, the effect of stochasticity in intracellular reactions depends on the cell size and the number of DNA molecules that exist in the cell. Thus, if we consider two cells, one of which having undergone DNA duplication and being bigger than the other, stochasticity in intracellular reactions will result in different phenotypic variations for these cells. Consequently, a mathematical framework that can capture the effects of the interactions of the several sources of heterogeneity may reveal novel effects that could not be revealed by studying these sources in isolation.

Furthermore, the necessity of focusing on populations, rather than on single cells,
becomes evident when the number of cells in the population may change as a result of the function of the genetic network under consideration. Typical examples of such genetic networks are those that contribute to the survival of a cell population under environmental stress conditions (heat shock protein systems, antibiotic resistance etc.), post-segregational killing systems, or the cell cycle. Attempting to simulate such systems focusing on the single cell will yield no information about the size of the population. Moreover, it is important to take into account the dynamics of the whole cell population when it splits into subpopulations of different phenotypes, especially when such phenotypes have differential fitness. In the case, for example, where two phenotypes exist with one phenotype growing faster that the other, one expects that the faster growing phenotype will become dominant. Now, suppose that a drug is administered in the environment and only the slower growing phenotype exhibits drug resistance. In this case it is not trivial to predict which phenotypes and at what fractions will be observed in the cell population. Such viability studies could give tremendous insight in patient recovery and relapse during treatment of a disease with a drug.

We have thus argued for the necessity of developing a generic mathematical formulation that can incorporate the major sources of stochasticity at the single cell and the cell population level. To this end, in Chapter 2 we develop a M-Equation and a corresponding MC algorithm that pertains to single cell processes and accounts for stochasticity in intracellular reactions, DNA duplication and cellular division. In Chapter 3 we formulate a M-Equation that takes into account these processes in the cell population level and derive the corresponding CPB by averaging. We further develop a MC algorithm that can be used to simulate the cell population M-Equation. In the second
part of this thesis, we apply these ideas to elucidate how stochasticity shapes single cell behavior and cell population heterogeneity for the case of an artificial lac operon genetic network, characteristic of positive feedback architecture. Thus, we first review the experimental and modeling literature for the lac operon system (Chapter 4). We then present deterministic and stochastic simulation at the single cell level, thereby demonstrating how stochasticity generates complex behavior (Chapter 5), and we subsequently turn our attention to the cell population level where we simulate the emergence of phenotypic heterogeneity (Chapter 6). Finally, in Chapter 7, we propose further work that can improve and extend the framework and review biological problems for which the approach developed in this thesis can give significant insight.
Chapter 2:

Modeling the Major Sources of Stochasticity in Intracellular Processes at the Single Cell Level

In this chapter we develop the single cell mathematical framework that takes into account all the sources of isogenic heterogeneity discussed in the introduction.

2.1 Preliminary Definitions

In this work, we will refer to the set of all the daughter cells of a mother cell as the cell's lineage (Figure 2.1). A cell population consists of all the lineages of the cells initially present in our observation-simulation. A line of cells is a collection of cells defined by the following recursive procedure: start from one mother cell; choose one of its daughter cells; set the daughter to mother cell; repeat (Figure 2.1). The objective of this chapter is to formulate an equation that refers to a cell line. Note that, when observing the history of a cell line, we are always referring to a single cell.

In accordance with previous work on the population balance framework (Ramkrishna 2000), we assume that each cell can be completely described by a state vector that contains information about the chemical content of the cell and the morphometric
Figure 2.1: Explanation of the terms lineage and line as used in this work. Lineage is the set of all the daughter cells of a mother cell. Line is a collection daughter cells defined recursively: start from one mother cell; choose one of its daughter cells; set the daughter to mother cell; repeat.

characteristics such as the length, the membrane area or the volume. For our work we will utilize only one morphometric characteristic, the volume. Assuming that the cells have the same shape but not the same size, all other morphometric characteristics can be computed if the volume is known. Thus, the state vector of the cell is an $N + 1$ sized vector ($N$ entries for species copy numbers and 1 entry for the volume). Let this vector be denoted as $\mathbf{z}$:

$$\mathbf{z} = (\mathbf{X}, V) \in \mathcal{Z} = (\mathbb{N}_0)^n \times \mathbb{R}_0^+$$

(2.1)

2.2 Sources of Stochasticity in Intracellular Processes

Due to stochastic effects in the intracellular processes, the state vector $\mathbf{z}$ of the cell under observation wanders randomly around $\mathcal{Z}$. Therefore, we need to construct a $M$-Equation for the probability of finding $\mathbf{z}$ at some point in the space as time passes. To
further clarify the nature of this M-Equation, let us use the concept of an ensemble.

Consider a large ensemble of independent cell lines (note that this ensemble is not a cell population). As time passes the states $z_1, z_2, \ldots$ of the cells in the ensemble change. Thus, suppose that we focus at a specific neighborhood of $Z$, and we count how many cells have state vectors entering or exiting the neighborhood. This inflow-outflow is loosely speaking equivalent to the inflow-outflow of probability in $Z$. Further, the M-Equation sought is a balance that equates the change in the probability content of the neighborhood in discussion, with the net probability flux in and out of the neighborhood:

$$P(X, V, t + \Delta t) \cdot \Delta V - P(X, V, t) \cdot \Delta V = \sum_{k=1}^{4} \text{ProbIn}_k - \sum_{k=1}^{4} \text{ProbOut}_k$$  \hspace{1cm} (2.2)

The net probability flux is a sum of components each of which pertains to a specific intracellular process, namely chemical reactions, DNA duplication, cell growth, and division. In the following, we will discuss these processes and will formulate the terms expressing contribution of each of them to the overall probability flux.

2.2.1 Chemical Reactions

Chemical reactions can result in the production or degradation of molecules, the synthesis of a new molecule from other molecules that serve as building blocks or the fission of a molecule to its building blocks. Since we attempt to build a generic framework we need to utilize a generic formulation that will allow us to incorporate any chemical reaction network in the final M-Equation.

Thus, let $S$ be a vector with the chemical species of interest, having the following structure:
The total number of species is defined as:

$$N = n + \sum_{i=1}^{d} s_i$$  \hspace{1cm} (2.4)$$

where $n$ is the number of non-chromosomal DNA species and $d$ the number of chromosomal DNA species. The necessity for discriminating between chromosomal and non-chromosomal species comes from the fact that upon division, chromosomal DNA species are partitioned equally in the two daughters. However, this is not generally true for the other species. Furthermore, each of the chromosomal DNA species $i = 1, \ldots, d$ may exist in $s_i$ states, for example an operator may exist in three states: the free state $O$, the repressed state with one repressor molecule bound $RO$, or the repressed state with two repressor molecules bound $R_2O$. Thus, for this case, $s_1 = 3$ and $(S_{n+1}, S_{n+2}, S_{n+3}) = (O, RO, R_2O)$.

The chemical species of interest are assumed to interact according to a general chemical reaction network of $m$ reactions with the $N$ participating species $S_i$:

$$\left\{ \sum_{i=1}^{N} \alpha_{ij} S_i \xrightarrow{k_j} \sum_{i=1}^{N} \beta_{ij} S_i \right\}_{j=1}^{m}$$  \hspace{1cm} (2.5)$$

where $k_j$ is the deterministic reaction rate constant (intensive quantity). This essentially is the network of biochemical reactions that models the biological system or pathway of interest. In order to be able to assess the effect of reactions on the state of the cell we need to know how the species copy numbers change once a specific reaction event has occurred, and how frequently such reaction events occur.

To answer the first question, if $X_i$ denotes the copy number (number of molecules) of
species $S_i$, we can define a vector $v_j$ which expresses the change in the contents $X$ of the
cell as reaction $j$ occurs in a cell. This vector is given as:

$$v_j = \{b_{ij} - \alpha_{ij}\}_{i=1}^{N}$$

(2.6)

e.g. if the reaction is $A + B \rightarrow C$ and the species vector is $[A \ B \ C]$ then $v_j = [-1 \ -1 \ 1]^T$.

To determine how frequently reactions occur, we consider the propensity function for
reaction $j$, $a_j(X, V)$, which is the stochastic analogue of a reaction rate. The propensity
function gives the probability density that one reaction-$j$ event will happen in the $(t, t +
dt)$ time interval. Thus, the larger the propensity function of reaction $j$, the more likely it
is that many reaction events of index $j$ ($j = 1, \ldots, m$) will happen during a time interval.
Furthermore, the propensity that any reaction is going to happen is equal to the sum of
the propensities (since the reaction occurrence events are mutually exclusive):

$$a_r(X, V) = \sum_{j=1}^{m} a_j(X, V)$$

(2.7)

Particular expressions for the propensity functions will be given in a subsequent section
(2.4 Particular Expressions for the Propensity Functions) since here we are primarily
interested in deriving the generic M-Equation for the single cell processes into
consideration.

Now we are ready to derive the contribution of the reaction occurrences to the overall
probability flux from/to state $(X, V)$. Loosely speaking, $P(X-v_j, V, t)$ is the fraction of cells
in our ensemble that exist in state $(X-v_j, V)$ at time $t$. Of these cells, a fraction equal to
$a_j(X-v_j, V) \cdot \Delta t$ will undergo reaction $j$ during the infinitesimal time interval $(t, t + \Delta t)$.
Subsequently, this fraction will end up in state $(X, V)$. If we apply the argument to all
kinds of reactions, it follows that the inflow of probability to state $(X, V)$ due to reactions
is:

$$\text{ProbIn}_1 = \sum_{j=1}^{m} a_j \left( X - v_j, V \right) \cdot \Delta t \cdot P \left( X - v_j, V, t \right) \cdot \Delta V$$  \hspace{1cm} (2.8)$$

One can easily repeat the argument just discussed to show that the outflow of probability from state $(X,V)$ due to reaction is:

$$\text{ProbOut}_1 = \sum_{j=1}^{m} a_j \left( X, V \right) \cdot \Delta t \cdot P \left( X, V, t \right) \cdot \Delta V \hspace{1cm} (2.9)$$

### 2.2.2 DNA Duplication

During duplication the chromosomal DNA species are doubled. We assume that the newly produced chromosomal DNA species $i$ exists in a basal state $\eta_i$. For example, in the case of an operator existing in the free and the two bounded states, the basal state will be the free state. Then the production of the new DNA will schematically be:

$$\left\{ \sum_{j=0}^{s_i-1} S_{\eta_i+j} \right\}_{j=1}^{d} \rightarrow S_{\eta_i} \hspace{1cm} \text{where:} \hspace{0.5cm} \eta_i = n + \sum_{j=1}^{i-1} s_j + 1$$  \hspace{1cm} (2.10)$$

The change in the species copy numbers is given by vector $v_s$, of which all elements are equal to zero except those that correspond to each state $\eta_i$ of DNA species $i$:

$$v_s = \left\{ \sum_{j=1}^{d} \delta_{\eta_i,k} \right\}_{k=1}^{N} \hspace{1cm} (2.11)$$

Moreover, the DNA duplication propensity function is $a_s \left( X, V \right)$ and expresses the probability that duplication will happen in the next $\Delta t$ as a function of the cell’s state.

Thus, similarly to the case of the reactions, the probability influx and outflux due to duplication will be:

$$\text{ProbIn}_2 = a_s \left( X - v_s, V \right) \cdot \Delta t \cdot P \left( X - v_s, V, t \right) \cdot \Delta V$$  \hspace{1cm} (2.12)$$
One can easily repeat the argument just discussed to show that the outflow of probability from state \((X, V)\) due to reaction is:

\[
\text{ProbOut}_2 = a_i (X, V) \cdot \Delta t \cdot P(X, V, t) \cdot \Delta V
\]  

(2.13)

2.2.3 Growth

The aforementioned reactions are assumed to be taking place in the volume of the single cell, \(V(t)\), which is given by solving a differential equation expressing cell growth:

\[
\frac{dV}{dt} = g(X, V)
\]  

(2.14)

Note that growth in this formulation is assumed to be a deterministic process once \(X\) and \(V\) have been defined. Stochasticity comes from the randomness in the state of the cell.

Furthermore, \(g(X, V) \cdot \Delta t\) is the volume change that will occur during the infinitesimal time interval \((t, t + \Delta t)\). Now, \(P(X, V, t)\) is the fraction of cells in our ensemble that exist in state \((X, V^-)\) and will undergo volume change \(g(X, V) \cdot \Delta t\). Thus, they will end up in state \((X, V^+)\), therefore the inflow of probability due to growth is:

\[
\text{ProbIn}_3 = g(X, V) \cdot \Delta t \cdot P(X, V, t)
\]  

(2.15)

By applying the argument just presented to the state \((X, V + \Delta V, t)\) where the instantaneous growth rate of the cell is \(g(X, V + \Delta V)\), we find that the outflow of probability due to growth is:

\[
\text{ProbOut}_3 = g(X, V + \Delta V) \cdot \Delta t \cdot P(X, V + \Delta V, t)
\]  

(2.16)

2.2.4 Division

It is further assumed that the cell divides with a propensity \(a_d(X, V)\) that is a function of the cell’s state. The partitioning of the content of the mother cell to the two daughters
is random, and thus we need to define the partitioning PDF \( h(X_d, V_d | X_m, V_m) \) which gives the probability of a daughter cell having contents \( X_d \) and volume \( V_d \), given that the mother had contents \( X_m \) and volume \( V_m \). The contents and the volume of the other daughter will then be given as \( X_m - X_d \) and \( V_m - V_d \) respectively. Thus, for the mass and volume to be conserved the following must hold:

\[
h(X, V | X+k, V+\delta) = h(k, \delta | X+k, V+\delta)
\]

(2.17)

The partitioning PDF has been introduced in the Population Balance framework (Ramkrishna 2000), but for the state variables being continuous quantities. It is apparent that function \( h \) can be constructed so that it expresses any partitioning law, such as binomial partitioning for non-chromosomal DNA species and equal partitioning with randomized state for the chromosomal DNA species.

We are now ready to derive the probability influx and outflux contributions for the division. The fraction of cells that exist in state \((X+k, V+\delta)\) and will undergo division during the time interval \((t, t + \Delta t)\) is given as \( a_d(X+k, V+\delta) \cdot \Delta t \cdot p(X, V+\delta, t) \cdot \Delta \delta \). Now, given the partitioning mechanism, what is the fraction of cells that result in state \((X, V)\), which is essentially the probability influx to this state? Note the following complication: upon division the ensemble size is doubled and essentially we randomly choose one of the two daughters to monitor (and thus, the ensemble size is restored).

According to the definition of the partitioning PDF, the quantity \( h(X, V | X+k, V+\delta) \cdot \Delta V \) gives the fraction of the dividing mother cells in the ensemble that, given that they exist in state \((X+k, V+\delta)\), produce a daughter with state \((X, V)\). Now, since we only monitor one daughter cell, the probability influx to state \((X, V)\) is the probability of the following event: daughter cell number 1 will be born having state \((X, V)\) and we will choose to
monitor it, or that daughter cell number 2 will be born having state \((X, V)\) and we will choose to monitor it. This event is equivalent to the following: daughter cell number 1 will be born having state \((X, V)\) and we will choose to monitor it, or that daughter cell number 1 will be born having state \((k, 9)\) and we will not choose to monitor it. Thus, if daughter cell 1 is chosen with probability \(\varpi\), since choosing which cell to monitor is independent of partitioning (otherwise we are biasing the process), and by using relation (2.17):

\[
\text{ProbIn}_{\text{from } (X+k, V+9)}^{(X+k, V)} = a_d (X+k, V+9) \cdot \Delta t \cdot P (X+k, V+9, t) \cdot \Delta 9
\]

\[
\left( \varpi \cdot h(X, V | X+k, V+9) + (1 - \varpi) \cdot \frac{h(k, 9 | X, V)}{h(X+k, V+9 | X, V)} \right) =
\]

\[
a_d (X+k, V+9) \cdot \Delta t \cdot P (X+k, V+9, t) \cdot h(X, V | X+k, V+9) \cdot \Delta 9
\]

Since all cells with state \((X+k, V+9)\) contribute to the influx at state \((X, V)\), we finally derive the influx of probability to state \((X, V)\) due to division as:

\[
\text{ProbIn}_4 =
\sum_{k_i > 0} \sum_{k_2 > 0} \sum_{i=0}^{\infty} a_d (X+k, V+9) \cdot \Delta t \cdot h(X, V | X+k, V+9) \cdot P (X+k, V+9, t) \cdot \Delta Vd9
\]

(2.19)

From the above discussion it should be clear that the fraction of cells that exist in state \((X, V)\) and will undergo division during the infinitesimal time interval \((t, t+\Delta t)\) is given as \(a_d(X,V)\cdot\Delta t\cdot P(X,V,t)\cdot\Delta V\). Thus, the outflow of probability due to division is:

\[
\text{ProbOut}_4 = a_d (X, V) \cdot \Delta t \cdot P(X, V, t) \cdot \Delta V
\]

(2.20)

### 2.3 The Single Cell M-Equation

In order to derive the M-Equation for the probability of finding a single cell in state \(z = (X,V)\) at time \(t\) we first substitute the probability influx and outflux terms of equations
(2.8, 9), (2.12, 13), (2.15, 16), (2.19, 20) into expression (2.2). We subsequently divide by $\Delta V \cdot \Delta t$ and take limits as $\Delta V \to 0$ and $\Delta t \to 0$, finally obtaining:

$$\frac{\partial P(X, V, t)}{\partial t} = \sum_{j=1}^{m} \left[ a_j(X - v_j, V) \cdot P(X - v_j, V, t) - a_j(X, V) \cdot P(X, V, t) \right]$$

$$+ a_s(X - v_s, V) \cdot P(X - v_s, V, t) - a_s(X, V) \cdot P(X, V, t)$$

$$- \frac{\partial}{\partial V} \left[ g(X, V) \cdot P(X, V, t) \right]$$

$$+ \sum_{k_1 \geq 0} \sum_{k_2 \geq 0} \int a_d(X + k_1, V + \theta) \cdot h(X, V | X + k_1, V + \theta) \cdot P(X + k_1, V + \theta, t) d\theta$$

$$- a_d(X, V) \cdot P(X, V, t)$$

(2.21)

Note that for the propensity functions to be physically meaningful, they must be equal to zero for any negative arguments. Similarly, the growth function must be zero for a cell of non-positive size (volume) or content:

$$a_j(X, V) = 0$$
$$a_s(X, V) = 0$$
$$a_d(X, V) = 0 \quad \text{if } \exists \ i : X_i < 0 \text{ or } V < 0$$
$$g(X, V) = 0$$

(2.22)

These restrictions prevent probability inflows (outflows) from (towards) quadrants of the space where content or volume takes negative values. Thus, we do not need to impose any special Boundary Conditions (BCs) on the formulation of the M-Equation. In the subsequent discussion, we will assume that (2.22) holds, without explicitly adding terms that nullify the propensities and growth term for negative arguments.

### 2.4 Particular Expressions for the Propensity Functions

In the previous sections, we derived the Single Cell M-Equation considering generic
expressions for the propensity functions that pertain to transitional events such as reactions and divisions. In this section we are going to discuss the particular forms of the propensity functions that will be used for this work.

2.4.1 Chemical Reactions

Consider a reaction network of the form (2.5) where chemical species interact inside a cellular volume \( V \) which is assumed to be well stirred. Following Gillespie (Gillespie 1976), we define the propensity function of reaction \( j \), \( a_j(X, V) \), as a function of the number of molecules \( X_i \) of species \( i \) and the volume of the "container" that hosts the interacting molecules:

\[
a_j(X, V) = k_j \cdot N_A \cdot V \cdot \prod_{i=1}^{N} \left[ \frac{\alpha_{ij}!}{(N_A \cdot V)^{\alpha_{ij}}} \left( \frac{X_i}{\alpha_{ij}} \right) \right]
\]

(2.23)

where \( N_A \) is Avogadro's number.

In order to model free diffusion of molecules through the membrane of the cell, we introduce zero-order reactions whose rate depends on the membrane surface area. Assuming high concentration and no depletion of the extracellular species into consideration, the propensity of transport of that species into the cytosolic space will be given as:

\[
a_j(X, V) = k_j \cdot A \cdot [S_e]
\]

(2.24)

where \([S_e]\) is the constant concentration of the extracellular species \( S_e \). Now we need to relate the area of the membrane of the cell the volume which is the only morphometric
characteristic that we consider. For this, we consider a rod-shaped bacterium having the length to width ratio shown in Figure 2.2. Then, it is easy to calculate that:

\[
V = \frac{13}{3} \cdot \pi \cdot R^3
\]

\[
A = 10 \cdot \pi \cdot R^2 = \frac{10}{13} \cdot 13^3 \cdot 3^3 \cdot \pi^3 \cdot V^3
\]

Therefore, the propensity function (2.23) can be evaluated by knowing the volume of the cell. Note that we have made the simplification of constant length to width ratio, and same for all the cells of the population.

If we wanted to consider a rod shaped bacterium with fixed diameter equal to \(\pi \cdot R_0^2\) but with length \(L\) that increases as the cell grows then the corresponding expressions for the volume and area would be:

\[
V = \frac{4}{3} \cdot \pi \cdot R_0^3 + \pi \cdot R_0^2 \cdot L
\]

\[
A = 4 \cdot \pi \cdot R_0^2 + 2 \cdot \pi \cdot R_0 \cdot L = \frac{2}{R_0} \cdot V + \frac{4}{3} \cdot \pi \cdot R_0^2
\]

The length to width ration would no longer be constant, in this case.

2.4.2 Growth

For our study, cell growth is assumed to be exponential following Cooper (1988). Other researchers have proposed liner, bilinear or other laws for the cell growth during one cell cycle. It is apparent that irrespectively of the particular growth law the model will always be able to reproduce Malthusian growth for the overall cell population.

However, we choose the exponential law (Cooper (1988)):

\[
\frac{dV}{dt} = g \cdot V
\]
Thus, given the state of the cell at time $t$ and assuming that this state does not change, the cell volume at times $t + \tau$ can be found for every positive $t$.

$$V(t + \tau) = \Phi(X, V(t), \tau) = V(t) \cdot e^{\xi \tau} \tag{2.28}$$

2.4.3 DNA Duplication

The DNA duplication propensity function is $a_\phi(X, V)$ and expresses the stochastic rate of the occurrence of duplication events as a function of the cell’s state. We use a volume dependent expression for the duplication propensity:

$$a_\phi(X, V) = \left(\frac{V}{V_{s, crit}}\right)^{n_1} \cdot \delta_{n=0} \sum_{j=n+1} X_j, U_1 \tag{2.29}$$

where the Kronecker delta $\delta$ is unity when the copy numbers of chromosomal species $1$ in any state sum to a nominal pre-duplication copy number $U_1$. This ensures that duplication is performed only once per cycle, when the chromosomal DNA species have copy number equal to $U$ and the cell volume (size) is close to $V_{s, crit}$.

2.4.4 Division

For the division propensity we use an expression similar to that used for DNA duplication:

$$a_d(X, V) = \left(\frac{V}{V_{d, crit}}\right)^{n_d} \cdot \delta_{n=1} \sum_{j=n+1} X_j, 2U_1 \tag{2.30}$$

Here we require the copy numbers of chromosomal DNA species to be equal to $2\cdot U$. The partitioning of the content of the mother cell to the two daughters is random, and thus we need to define the partitioning PDF $h(X_d, V_d | X_m, V_m)$, which gives the
probability of a daughter cell having contents $X_d$ and volume $V_d$ given that the mother had contents $X_m$ and volume $V_m$. The contents and the volume of the other daughter will then be given as $X_m - X_d$ and $V_m - V_d$, respectively. In order to construct a partitioning PDF we assume that volume partitioning is independent of content partitioning, and thus we can factorize the partitioning PDF to a term for volume partitioning, and a term for content partitioning. Following Ramkrishna (2000), the former term is assumed to have the form of a symmetric beta distribution:

$$
\beta(V_d \mid V_m) = \frac{1}{V_m} \frac{\Gamma(2 \cdot q)}{\left( \Gamma(q) \right)^2} \left( \frac{V_d}{V_m} \right)^{q-1} \left( 1 - \frac{V_d}{V_m} \right)^{q-1}
$$

(2.31)

where $q$ is a parameter controlling the sharpness of the division mechanism. Higher values of $q$ result in equal partitioning events being more probable.

We assume binomial partitioning for all non-chromosomal DNA species. The binomial partitioning of each species is performed independently and the “success probability” is equal to the daughter to mother volume ratio. This choice has the following physical meaning: during division, the mother cell “donates” each of the molecules it contains to one of the daughter cells. For each “donation event”, the probability that one molecule will result in the first daughter cell is $V_{d,i}/V_m$ (probability of success in each Bernoulli trial). Thus, the probability of the first daughter inheriting $X_{d,i}$ molecules of species $i$, given that the mother has $X_{m,i}$ molecules and the volumes of the daughter and mother are $V_d$ and $V_m$ respectively, is:

$$
b_i(X_{d,i} \mid X_{m,i}, V_m, V_d) = \binom{X_{m,i}}{X_{d,i}} \left( \frac{V_d}{V_m} \right)^{X_{d,i}} \left( 1 - \frac{V_d}{V_m} \right)^{X_{m,i} - X_{d,i}} \quad \text{for } i = 1, \ldots, n
$$

(2.32)

The chromosomal DNA species require symmetric partitioning (each daughter will
inherit equal DNA content) but with randomized state. Let us focus for example in chromosomal DNA species i, which may be an operator that may exist in the following three states: free (O), bounded with one repressor (RO), bounded with two repressors (R2O). The copy numbers of that species in the different states are given by vector $X_{m}^{DNA,i}$ (subscript m stands for "mother") defined as:

$$X_{m}^{DNA,i} = \{X_{m,k}\}^{n+\sum_{j}^{s_{j}}}_{k=0}^{}$$  \hspace{1cm} (2.33)

Also, let us set the copy number of DNA species i in any state for the mother and the daughter as:

$$M_{i} = \sum_{j=1}^{s_{i}} X_{m,j}^{DNA,i}$$  \hspace{1cm} (2.34)

$$D_{i} = \sum_{j=1}^{s_{i}} X_{d,j}^{DNA,i} = \frac{M_{i}}{2}$$  \hspace{1cm} (2.35)

Note that $M_{i} = 2U_{i}$ and $D_{i} = U_{i}$ (see also equations 2.29, 2.30 and pertinent text).

Then, if the mother cell has one operator in the bounded state and the other in the free state, the daughter cell may inherit any one of the two operators. Let us focus on DNA species i that may exist in $s_{i}$ states. From the theory of enumerative combinatorics we can calculate the probability of the daughter cell inheriting $X_{d,1}^{DNA,i}$ molecules at state 1 out of the $X_{m,1}^{DNA,i}$ that the mother has and $X_{d,2}^{DNA,i}$ out of the $X_{m,2}^{DNA,i}$ etc. This probability will be the product for all states $j$, of the combinations of $X_{m,j}^{DNA,i}$ per $X_{d,j}^{DNA,i}$ divided by the overall combinations of the total molecules of DNA species i in the mother per those in the daughter.
\[ c_i \left( X_d^{\text{DNA},i} | X_m, V_m, V_d \right) = \prod_{j=1}^{s_i} \begin{pmatrix} X_m^{\text{DNA},j} \\ X_d^{\text{DNA},j} \\ M_i \\ D_i \end{pmatrix} \] for \( i = 1, \ldots, d \) (2.36)

Note that this problem is similar to having \( 2n \) balls colored with \( m \) colors inside an urn and picking \( n \) balls out of the urn. We seek the probability of picking \( n_1 \) balls of the first color, \( n_2 \) of the second etc. The colors are equivalent to the states of the DNA species in our problem, and the balls are the chromosomal-DNA molecules.

Equation (2.36) can be transformed into an expression that will be more convenient for application in a MC simulation scheme. Let us refer to DNA species \( i \) and let the mother cell have \( X_m^{\text{DNA},i} \) copies of species \( i \) in each state \( j, j = 1, \ldots, s_i \). Similarly, the daughter cell will have \( X_d^{\text{DNA},i} \) copies of species \( i \) in each state \( j, j = 1, \ldots, s_i \). What is the probability that the daughter cell will inherit exactly \( X_d^{\text{DNA},i} \) copies out of the \( X_m^{\text{DNA},i} \) of the state 1? This probability will be the product of three terms: (i) the combinations of \( D_i \) per \( X_d^{\text{DNA},i} \), because any \( X_d^{\text{DNA},i} \) of the \( D_i \) molecules to be inherited may be in state 1; (ii) the combinations of \( M_i - D_i \) per \( X_m^{\text{DNA},i} - X_d^{\text{DNA},i} \), since any \( X_m^{\text{DNA},i} - X_d^{\text{DNA},i} \) of the \( M_i - D_i \) that will not be inherited must also be in state 1 so that exactly \( X_d^{\text{DNA},i} \) copies are inherited; (iii) the inverse of the total combinations which will be the combinations of \( M_i \) molecules per \( X_m^{\text{DNA},i} \) those in state 1. Now we can continue with the probability that the daughter cell will inherit exactly \( X_d^{\text{DNA},i} \) copies out of the \( X_m^{\text{DNA},i} \) of the state 2. This probability will again be the product of three terms: (i) the combinations of \( D_i - X_d^{\text{DNA},i} \) per \( X_d^{\text{DNA},i} \), because any \( X_d^{\text{DNA},i} \) of the remaining \( D_i - X_d^{\text{DNA},i} \) molecules to be inherited
may be in state 2; (ii) the combinations of \( M - D - (X_{m,1}^{\text{DNA,i}} - X_{d,1}^{\text{DNA,i}}) \) per \( X_{m,2}^{\text{DNA,i}} - X_{d,2}^{\text{DNA,i}} \),

since any \( X_{m,2}^{\text{DNA,i}} - X_{d,2}^{\text{DNA,i}} \) of the remaining \( M - D - (X_{m,1}^{\text{DNA,i}} - X_{d,1}^{\text{DNA,i}}) \) that will not be inherited must also be in state 2 so that exactly \( X_{d,2}^{\text{DNA,i}} \) copies are inherited; (iii) the inverse of the total combinations which will be the combinations of \( M - X_{m,j}^{\text{DNA,i}} \) molecules per \( X_{m,2}^{\text{DNA,i}} \) those in state 2. We can continue similarly to derive the probability of inheritance of exactly \( X_{d,j}^{\text{DNA,i}} \) copies out of the \( X_{m,j}^{\text{DNA,i}} \) of any subsequent state \( j \).

Note that by using the above arguments, we have essentially broken the process of partitioning to independent events of inheritance of \( X_{d,j}^{\text{DNA,i}} \) out of the \( X_{m,j}^{\text{DNA,i}} \). The independence allows us to express the overall probability of inheritance of \( X_{d}^{\text{DNA,i}} \) molecules out of the \( X_{m} \) as follows:

\[
\begin{align*}
 c_i \left( X_{d}^{\text{DNA,i}} \mid X_{m}, V_{m}, V_{d} \right) &= \prod_{j=1}^{s_i} \frac{D_i - \sum_{k=1}^{j-1} X_{d,k}^{\text{DNA,i}}}{X_{d,j}^{\text{DNA,i}}} \cdot \frac{M_i - D_i - \sum_{k=1}^{j-1} \left( X_{m,k}^{\text{DNA,i}} - X_{d,k}^{\text{DNA,i}} \right)}{X_{m,j}^{\text{DNA,i}} - X_{d,j}^{\text{DNA,i}}} \\
 &= \left( \sum_{k=j}^{s_i} X_{m,k}^{\text{DNA,i}} \right) / \left( \sum_{k=j}^{s_i} X_{m,j}^{\text{DNA,i}} \right) \\
 &= \frac{s_i}{\sum_{k=j}^{s_i} X_{m,k}^{\text{DNA,i}}} \\
 &= \frac{s_i}{\sum_{k=j}^{s_i} X_{m,j}^{\text{DNA,i}}}
\end{align*}
\]  

(2.37)

Finally, assuming that the partitioning occurs independently for every species (namely non-chromosomal DNA and chromosomal in any state), the overall partitioning PDF

\[
h(X_{d}, V_{d} \mid X_{m}, V_{m}) \] will be:

\[
h(X_{d}, V_{d} \mid X_{m}, V_{m}) = \\
\beta(V_{d} \mid V_{m}) \cdot \prod_{i=1}^{n} b_i(X_{d,i} \mid X_{m}, V_{m}, V_{d}) \cdot \prod_{i=1}^{d} c_i(X_{d}^{\text{DNA,i}} \mid X_{m}, V_{m}, V_{d})
\]  

(2.38)
2.5 MC Simulation of the Single Cell M-Equation

The Single Cell M-Equation is difficult to solve analytically or even numerically, if the chemical reaction network involves many species. Thus, we need to develop a MC simulation algorithm which will simulate reaction and division events, under the condition of continuous cell growth. From the previous description of the processes involved, one realizes that the random numbers of interest are the following:

1. **Pertaining to chemical reaction events:**

   \( \tau_{\text{ran}} \): the inter-arrival time for reaction events, namely the time between the last event simulated and the upcoming event of occurrence of any kind of reaction.

   \( \mu \): the index of reaction to be simulated \( \mu = [1, 2, \ldots, m] \) where \( m \) is the number of reactions (refer to equation 2.5)

2. **Pertaining to DNA duplication events:**

   \( \tau_{\text{dup}} \): the inter-arrival time for reaction events, namely the time between the last event simulated and the upcoming event of occurrence of duplication.

3. **Pertaining to division events:**

   \( \tau_{\text{div}} \): the inter-arrival time for division events, namely the time between the last event simulated and the upcoming division event.

   \( \rho \): the volume ratio of the mother to the daughter cell.

   \( \xi_i \): the number of molecules of species \( i \) that the first daughter cell will inherit.

The MC simulation will proceed by computing the times of the occurrence of the next reaction, duplication and division event. Then, the event with the smallest waiting time will be simulated by drawing additional random numbers if necessary. Thus, in order to completely define the MC simulation we need to define the probability mass or density
functions for the above random numbers.

2.5.1 Inter-arrival times for reaction events

For the derivations of the inter-arrival time distributions we make use of the concept of the interval of quiescence (Shah et al. 1977). The event that some reaction will happen after a waiting time \( \tau_r \) is the intersection of the events (i) that no reaction will happen in the interval \( (t, t + \tau_r) \) and (ii) that some reaction will happen at time \( t + \tau_r \). These events are independent. Thus, the probability density of some reaction (of any type) happening after waiting time \( \tau_r \) is the product of the probability of no reaction event happening in the interval \( (t, t + \tau_r) \) and the probability density of any reaction event happening exactly at time \( t + \tau_r \). Let us first consider the probability of occurrence of no reaction event, given that the cell exists in state \((X,V)\) at time \( t \). The following notation will be used:

\[
p_{\text{no rxn}} \left( \text{ending time}\,|\,\text{initial time,}\,\begin{cases} \text{initial number of molecules} \\
\text{volume through time interval} \end{cases} \right)
\]

Note that the initial number of molecules stays constant throughout the time interval that no reaction happens. Apparently, the probability that no reaction will happen at time \( t = 0 \) is equal to 1; this information will be used as an IC. Furthermore, since \( a_r(X,V) \) is the probability density that some reaction is going to happen at the next \( d\tau_r \) time interval we can write the following probability balance:

\[
p_{\text{no rxn}} \left( t + \tau_r + d\tau_r \bigg| t, X, \Phi(X, V(t), \tau_r + d\tau_r) \right) = \\
p_{\text{no rxn}} \left( t, t + \tau_r \bigg| X, \Phi(X, V(t), \tau_r) \right) \cdot \left( 1 - a_r \left( X, \Phi(X, V(t), \tau_r) \right) \right) \\
\cdot d\tau_r
\]

subject to:

\[
p_{\text{no rxn}} \left( t \bigg| t, X, V(t) \right) = 1
\]

For the definition of \( \Phi \) see equation (2.28). Therefore:
\[
\frac{d}{d\tau_r} \left[ \ln \left( p_{so,\text{rn}}(t + \tau_r, t, X, \Phi(X, V(t), \tau_r + d\tau_r)) \right) \right] = -a_r(X, \Phi(X, V(t), \tau))
\Rightarrow
\]
\[
p_{so,\text{rn}}(t + \tau_r, t, X, \Phi(X, V(t), \tau_r)) = \exp \left[ -\int_0^{\tau_r} a_r(X, \Phi(X, V(t), \tau')) d\tau' \right]
\]

(2.41)

Now, the probability density that any reaction event will happen exactly at time \( t + \tau_r \) is \( a_r(X, \Phi(V(t), \tau)) \). Therefore, the probability that the first reaction (of any kind) after time \( t \) will happen at time \( t + \tau_r \) is:

\[
p_{\text{some, rn}}(t + \tau_r, t, X, \Phi(X, V(t), \tau_r)) = \\
a_r(X, \Phi(X, V(t), \tau_r)) \cdot \exp \left[ -\int_0^{\tau_r} a_r(X, \Phi(X, V(t), \tau')) d\tau' \right]
\]

(2.42)

Equation (2.42) defines the probability density of the inter-arrival times (they can also be called waiting times) of reaction events. This equation needs the following information: the current state of the cell, \((X, V(t))\), the cellular growth expression \( \Phi \) that essentially gives the volume throughout the waiting-time interval, and finally the propensity functions of each reaction. The state of the cell is known at each step of the MC algorithm and the cellular growth expression \( \Phi \) is obtained by solving differential equation (2.27). Thus, one can generate random numbers following density (2.42) at any stage of the MC run (see Appendix I for how to map a uniform deviate to a random number following a desired distribution).

Note that the probability density (2.42) is not necessarily normalized to unity. In other words, there may be cases where there is a finite probability that no reaction occurs in the future. This probability can be calculated as follows:
\[
\lim_{\tau_r \to \infty} p_{\text{on reaction}}(t + \tau_r | t, X, \Phi(X, V(t), \tau_r)) = \exp\left[-\int_0^{\infty} a_r(X, \Phi(X, V(t), \tau'))d\tau'\right]
\] (2.43)

Therefore, if the integral \( I_\infty \) diverges to infinity, it is almost sure that at least one reaction is going to occur in finite time. From the particular functional forms of the propensity functions, and assuming that \( \Phi \) is monotonically increasing, one can easily deduce that for a network containing 0th and 1st order reactions \( I_\infty \) diverges to infinity.

2.5.2 Index of reaction to be simulated

Once the MC algorithm has determined that a reaction event is going to occur (for this to be true, \( I_\infty \) must diverge to infinity in equation 2.43), then a random number needs to be generated in order to determine which reaction is going to take place. This random number \( \mu \) follows the probability mass function (Gillespie 1977; Lu et al. 2004):

\[
P_{\text{reaction index } = \mu} = \frac{a_\mu(X, \Phi(X, V(t), \tau_r))}{\sum_{k=1}^m a_k(X, \Phi(X, V(t), \tau_r))}
\] (2.44)

Thus, given the propensity functions at time \( t + \tau_r \), the algorithm can generate a random number that represents which reaction event must be simulated (see Appendix I for how to map a uniform deviate to a random number following a desired distribution).

2.5.3 Inter-arrival times for duplication events

Similarly to the inter-arrival times of reaction events, one can derive the probability density for the DNA duplication events. The result is:
\[ P_{\text{DNA dupl}}(t + \tau_s|t, \mathbf{X}, \Phi(\mathbf{X}, V(t), \tau_s)) = \]
\[ a_s(\mathbf{X}, \Phi(\mathbf{X}, V(t), \tau_s)) \cdot \exp \left[ - \int_0^{\tau_s} a_s(\mathbf{X}, \Phi(\mathbf{X}, V(t'), \tau')) \, dt' \right] \]  
(2.45)

The probability density \(a_s(\mathbf{X}, V)\) has to be chosen carefully, so that the cell divides after some random time that is distributed around some fraction of the \textit{E. coli} division time (the latter is 25 - 45 min for \textit{E. coli} cells). To avoid infinite duplication times (equivalently: no future duplication events) the integral of \(a_s\) has to diverge to infinity:
\[ \int_0^{\infty} a_s(\mathbf{X}, \Phi(\mathbf{X}, V(t), \tau')) \, dt' \rightarrow \infty \]  
(2.46)

2.5.4 Inter-arrival time for division events

Similarly to the inter-arrival times for reaction or DNA duplication events, the inter-arrival times for division events will be given as:
\[ P_{\text{no div}}(t + \tau_d|\mathbf{X}, \Phi(\mathbf{X}, V(t), \tau_d)) = \]
\[ a_d(\mathbf{X}, \Phi(\mathbf{X}, V(t), \tau_d)) \cdot \exp \left[ \int_0^{\tau_d} a_d(\mathbf{X}, \Phi(\mathbf{X}, V(t'), \tau')) \, dt' \right] \]  
(2.47)

To avoid infinite division times:
\[ \int_0^{\infty} a_d(\mathbf{X}, \Phi(\mathbf{X}, V(t), \tau')) \, dt' \rightarrow \infty \]  
(2.48)

2.5.5 Volume ratio of the mother to the daughter cell

Once the MC algorithm has determined that a division event is going to occur, then a random number must be generated in order to determine the volume ratio of mother to daughter cell. This random number \(\rho \in [0,1]\) must follow a symmetric distribution in the sense that \(\rho\) and \(1 - \rho\) must be identically distributed. In accordance with equation (2.31):
\[ \rho \sim \beta(q, q) \]  

(2.49)

2.5.6 Number of non-chromosomal molecules inherited by one daughter cell

For partitioning the non-chromosomal DNA species, the algorithm has to determine for each species how many molecules will be inherited by one daughter. Thus, \( n \) random numbers have to be generated: one for each species. In accordance with (2.32) these numbers will be chosen to follow binomial distributions with probability of success equal to \( \rho \) and number of trials equal to the molecular content of the mother cell.

\[ \xi_i \sim \delta(\rho, X_{m,i}) \quad \text{for } i = 1, \ldots, n \]  

(2.50)

2.5.7 Number of chromosomal molecules inherited by one daughter cell

For partitioning the chromosomal DNA species, the algorithm will apply equal partitioning on the number of molecules but it will still have to determine the states of the inherited DNA molecules. In accordance to the discussion of section 2.2.4, for the algorithm to determine how many molecules of chromosomal DNA species \( i \) will be inherited in state \( j \), it suffices to generate a random number \( \xi_{\sum_{k=1}^i n_s + j} \) that will follow a hypergeometric distribution denoted noted as \( hg(#Total, #Defective, #Draws) \):

\[ \xi_{\sum_{k=1}^i n_s + j} \sim hg \left( \sum_{k=j}^{s_i} X_{m,k}^{DNA,i}, D_j - \sum_{k=1}^{j-1} \xi_{\sum_{k=1}^i n_s + k}, X_{m,j}^{DNA,i} \right) \]  

for \( j = 1, \ldots, s_i \)  

for \( i = 1, \ldots, d \)  

(2.51)

Note that for the generation of each random number pertaining to state \( j \) the quantities that enter the calculation are always known: they are either mother cell contents or daughter cell contents of states \( 1, \ldots, j - 1 \).
2.5.8 Pseudo-code

Given the reaction network, the growth function $g$, the propensities $a_s, a_d$, and the partitioning mechanism $h$, the procedures to be simulated are as follows.

1. **Begin execution**

2. Define the initial and final times of the simulation. Initialize time as $t = \text{initial time}$

3. Define the initial molecular contents $X$ and volume $V$ of the cell

4. Define a constant $\Delta t_{sp}$ for sampling through time and define the time that the first sample will be taken $t_{sp} = \text{initial time}$

5. Open a file `filename` for writing time-course data

6. Do while $t < \text{final time}$:

   6.1. Calculate the waiting time for the next reaction event, $\tau_{rxn}$
   
   6.2. Calculate the waiting time for the next duplication event, $\tau_{dup}$
   
   6.3. Calculate the waiting time for the next division event, $\tau_{div}$
   
6.4. If $\tau_{rxn} < \tau_{dup}$ and $\tau_{rxn} \leq \tau_{div}$ then a reaction event will happen:

   6.4.1. Find $\mu$ the kind of reaction to be simulated
   
   6.4.2. Store the previous state $X_{old} = X, V_{old} = V$
   
   6.4.3. Perform the reaction, update the state $X, V$
   
   6.4.4. Update time $t = t + \tau_{rxn}$

6.5. Else if $\tau_{dup} \leq \tau_{rxn}$ and $\tau_{dup} < \tau_{div}$ then a DNA duplication event will happen:

   6.5.1. Perform the duplication, update the state $X, V$
   
   6.5.2. Update time $t = t + \tau_{rxn}$

6.6. Else if $\tau_{div} \leq \tau_{dup}$ and $\tau_{div} < \tau_{rxn}$ and then a division event will happen:

   6.6.1. Find $\rho$ the ratio of the volumes of the first daughter to the mother cell
6.6.2. For every species i find $\xi_i$, the number of molecules of species i that the first daughter cell will inherit

6.6.3. Store the previous state $X_{old} = X, V_{old} = V$

6.6.4. Perform the division, update the state $X, V$

6.6.5. Update time $t = t + \tau_d$

6.7. Sampling in time: Do while $t_{sp} < t$:

6.7.1. write into filename1 data $t_{sp}, X_{old}, V$

6.7.2. Update sampling time: $t_{sp} = t_{sp} + \Delta t_{sp}$

7. Close the file filename1

8. Terminate execution

2.6 A Simple Model-Problem

In the previous sections we developed a M-Equation for the following single cell processes: cellular growth, intracellular reactions, DNA duplication and cellular division.

In order to illustrate the processes involved and test the algorithm we will use a simple network of reactions involving only one species and two reactions: a $0^{th}$ order production and a $1^{st}$ order degradation. The model-problem was chosen to only involve one species so that we can numerically solve the M-Equation and compare the results with those obtained from MC simulations. Thus, we can show convergence in the mean square sense.

2.6.1 Reaction Network, Growth and Division Mechanism

The chemical reaction network be simulated under the conditions of cell growth and division consists of the following reactions:
\[ \varnothing \xrightarrow{k_1} X \]  (2.52)

\[ X \xrightarrow{k_2} \varnothing \]  (2.53)

where \( k_1 \) and \( k_2 \) are the "deterministic" intensive reaction rate constants. Therefore, from equation (2.23) the propensity functions are:

\[ a_1 (X, V) = k_1 \cdot N_A \cdot V \]  (2.54)

\[ a_2 (X, V) = k_2 \cdot X \]  (2.55)

where \( N_A \) is Avogadro’s number, \( X \) the number of molecules, and \( V \) the cell volume.

Single cell growth follows the exponential expression of equation (2.27):

\[ \frac{dV}{dt} = g \cdot V \]  (2.27)

For this example no DNA species exist, and therefore no DNA duplication was considered. Furthermore, the division propensity is taken as in equation (2.30) without the delta term:

\[ a_d (X, V) = \left( \frac{V}{V_{d, \text{crit}}} \right)^{n_d} \]  (2.56)

The partitioning PDF \( h(X_d, V_d | X_m, V_m) \), where the indexes \( m \) and \( d \) denote mother and daughter respectively, is the product of two terms, one referring to the volume and the other to the content of the only (non-DNA) species (refer to equation 2.38):

\[ h(X_d, V_d | X_m, V_m) = \beta (V_d | V_m) \cdot b(X_d | X_m, V_m, V_d) \]  (2.57)

The term for the volume is given by equation (2.31):

\[ \beta (V_d | V_m) = \frac{1}{V_m} \cdot \frac{\Gamma (2 \cdot q)}{(\Gamma (q))^2} \cdot \left( \frac{V_d}{V_m} \right)^{q-1} \cdot \left( 1 - \frac{V_d}{V_m} \right)^{q-1} \]  (2.31)

and the term for the content is given as (refer to equation 2.32):
\[
\begin{align*}
&b(X_d | X_m, V_m, V_d) = \left( \frac{X_m}{X_d} \right) \left( \frac{V_d}{V_m} \right)^{X_m} \left( 1 - \frac{V_d}{V_m} \right)^{X_m - X_d} \\
&\text{(2.58)}
\end{align*}
\]

Therefore the M-Equation for this system is:

\[
\begin{align*}
&\frac{\partial P(X, V, t)}{\partial t} = a_1 (X - 1, V) \cdot P(X - 1, V, t) - a_1 (X, V) \cdot P(X, V, t) \\
&\quad + a_2 (X + 1, V) \cdot P(X + 1, V, t) - a_2 (X, V) \cdot P(X, V, t) \\
&\quad - \frac{\partial}{\partial V} \left[ g \cdot V \cdot P(X, V, t) \right] \\
&\quad + \sum_{k \geq 0} a_d (X + k, V + 9) \cdot h(X, V | X + k, V + 9) \cdot P(X + k, V + 9, t) d9 \\
&\quad - a_d (X, V) \cdot P(X, V, t) \\
&\text{(2.59)}
\end{align*}
\]

The IC used was a delta function:

\[
P(X, V, t)|_{t=0} = \delta_{X, X_0} \cdot \delta (V - V_0) \\
\text{(2.60)}
\]

2.6.2 Numerical Solution of the M-Equation and MC Simulations

**Truncation of the State Space**

Equation (2.59) will be solved numerically within a subset of the state space:

\[
(X, V) \in \{0, ..., X_{\text{max}}\} \times [V_{\text{min}}, V_{\text{max}}] \\
\text{(2.61)}
\]

Note that (2.59) is written for the entire \( N_0 \times \mathbb{R}_0^+ \), and thus we impose a "truncation" of the state space. This is numerically acceptable, since most of the probability density is essentially concentrated in a small region of the state space. However, we will have to impose artificial BCs in order to contain our solution in this rectangle and avoid outflow of probability which will violate the normalization condition. These conditions were periodic BCs in the V direction and no flux conditions for \( X=X_{\text{max}} \) (note that we do not have to apply anything for \( X = 0 \) since 2.22 holds). Thus, we are essentially solving
equation (2.59) with the following BCs:

\[
\left. \frac{\partial}{\partial V} \left[ g \cdot V \cdot P(X, V, t) \right] \right|_{V=V_{\text{min}}} = \left. \frac{\partial}{\partial V} \left[ g \cdot V \cdot P(X, V, t) \right] \right|_{V=V_{\text{max}}} \tag{2.62}
\]

\[
\left. \frac{\partial P(X, V, t)}{\partial t} \right|_{X=X_{\text{max}}} = a_1 (X-1, V) \cdot P(X-1, V, t) - a_2 (X, V) \cdot P(X, V, t)
\]

\[
- \frac{\partial}{\partial V} \left[ g \cdot V \cdot P(X, V, t) \right]
\]

\[
+ \sum_{k \geq 0} a_k (X+k, V+\theta) \cdot h(X, V | X+k, V+\theta) \cdot P(X+k, V+\theta, t) d\theta
\]

\[- a_d (X, V) \cdot P(X, V, t) \tag{2.63}
\]

Needless to say, the rectangle where we solve the discretized equation numerically has to be large enough so that the numerically calculated probability is a good approximation of the true solution.

**Discretization**

The M-Equation was discretized using a 1st order finite difference scheme for the volume \( V \) and Forward Euler as a time integrator. The discretized spatial interval is:

\[
I_{ds} = \{0, \ldots, X_{\text{max}} \} \times \left\{ V_{\text{min}} + \left( V_{\text{max}} - V_{\text{min}} \right) \cdot \frac{i-1}{N_{\text{nodes}}-1}, \ i = 1, \ldots, N_{\text{nodes}} \right\} \tag{2.64}
\]

Apparentely the spatial stepsize \( \Delta V \) is:

\[
\Delta V = \frac{V_{\text{max}} - V_{\text{min}}}{N_{\text{nodes}} - 1} \tag{2.65}
\]

In turn, the discretized time interval is:

\[
I_{dt} = \{ \Delta t \cdot (j-1), \ j = 1, 2, \ldots \} \tag{2.66}
\]

The discretized equation for an internal node is:
\[
P_{X,i}^{(j+1)} = P_{X,i}^{(j)} + \Delta t \cdot \left\{ \begin{array}{l}
a_1 (X-1, V_i) \cdot P_{X,i}^{(j)} - a_i (X, V_i) \cdot P_{X,i}^{(j)} \\
+ a_2 (X+1, V_i) \cdot P_{X,i+1}^{(j)} - a_2 (X, V_i) \cdot P_{X,i}^{(j)} \\
- g \cdot \frac{1}{\Delta V} \left[ P_{X,i}^{(j)} - P_{X,i-1}^{(j)} \right] \\
+ \sum_{k=1}^{X_{\max}} \sum_{m=i}^{i_{\max}} a_d (X+k, V_m) \cdot h(X, V_i | X+k, V_m) \cdot P_{k,m}^{(j)} \\
- a_d (X, V_i) \cdot P_{X,i}^{(j)} \end{array} \right\} \tag{2.67}
\]

where \(P_{X,i}^{(j)} = P(X, V_{i\min} + i \cdot \Delta V, j \cdot \Delta t) = P(X, V_i, t_j)\) as \(\Delta V \to 0, \Delta t \to 0\). This equation has to be accompanied by the appropriate BCs. The discretized equation for BC (2.63) can be easily derived and will not be presented here. The discretized version of BC (2.62) can be written it as follows:

\[
- g \cdot \frac{1}{\Delta V} \left[ P_{X,i_{\max}}^{(i)} - P_{X,i_{\min}^{-1}}^{(i)} \right] = - g \cdot \frac{1}{\Delta V} \left[ P_{X,i_{\max}}^{(i)} - P_{X,i_{\max}^{-1}}^{(i)} \right] \tag{2.68}
\]

The "hypothetical node" \(i_{\min}^{-1}\) that appears as an index to \(P_{X,i_{\min}^{-1}}^{(i)}\) needs to be eliminated. Thus, if we solve for \(P_{X,i_{\min}^{-1}}^{(i)}\) it trivially follows that:

\[
P_{X,i_{\min}^{-1}}^{(i)} = P_{X,i_{\max}}^{(i)} \tag{2.69}
\]

and thus, we can substitute the value \(P_{X,i_{\max}}^{(i)}\) to the discretized equation (2.67) written for nodes \((X,i_{\min}) \forall X \epsilon \{0,...,X_{\max}\}\).

The discretized equation for the IC is written as:

\[
P_{X_0,0}^{(0)} = \frac{1}{\Delta V} \quad \text{where} \quad i_0 = \text{floor} \left( \frac{V_0}{\Delta V} \right) + 1 \tag{2.70}
\]

There are two problems that arise from the discretization just discussed:

(i) The partitioning function \(h(X, V_i | X+k, V_m)\) is a probability and has to be
normalized so that it integrates to unity. However, the discretization in the V
direction introduces truncation errors; as a result, the probability $P_{x,i}$ will not be
conserved as the Forward Euler scheme proceeds in time. As a remedy to this
problem, $h$ was pre-computed, and the result was renormalized so that it integrates
to unity.

(ii) A false diffusion effect is introduced due to the term $g \left[ P_{x,i}^{(j)} - P_{x,i-1}^{(j)} \right]/\Delta V$,
which models cell growth and is essentially a convection term. The problem of
false diffusion is attributed to the truncation error resulting from the discretization
of the differential equation. There is no false-diffusion-free scheme and false
diffusion becomes significant when applying the numerical scheme in convection
dominated processes. False diffusion is eliminated when the Courant number (in
our case equal to $g_i V_i \Delta t/\Delta V$) for node $i$ is equal to unity. Yet, Courant numbers
close to unity can create instabilities, and, in any case for our model-problem, we
cannot have uniform Courant number throughout the interval. However, false
diffusion in our case only affects the transient part of the solution. Once stationary
state has been reached, the accuracy of the solution depends on the $\Delta V$ and false
diffusion is no longer observed.

We thus applied the numerical scheme (2.67) to solve the M-Equation. The numerical
values for all the parameters of the system and the numerical schemes are shown in Table
2.1, where it is indicated whether a parameter has units of time, volume, etc. by using
actual units. The results appear in Figure 2.3: panel (a) shows the mean volume and
content as a function of time, as calculated from the probability profiles $P(X,V,t)$. For
large times both means approach a steady state, suggesting that the system reaches a time
Table 2.1: Parameters for the model problem and the simulators

<table>
<thead>
<tr>
<th>Parameters of the system</th>
<th>Parameters of the M-Equation solver</th>
</tr>
</thead>
<tbody>
<tr>
<td>$k_1$</td>
<td>40</td>
</tr>
<tr>
<td>$k_2$</td>
<td>$2 \cdot 10^{-5}$</td>
</tr>
<tr>
<td>$g$</td>
<td>1</td>
</tr>
<tr>
<td>$V_{d,\text{crit}}$</td>
<td>$10 \cdot 10^{-16}$</td>
</tr>
<tr>
<td>$n_d$</td>
<td>$2 \cdot 10^{16}$</td>
</tr>
<tr>
<td>$q$</td>
<td>80</td>
</tr>
<tr>
<td>$N_A$</td>
<td>$6.0221367 \cdot 10^{-14}$</td>
</tr>
<tr>
<td>$t_{\text{initial}}$</td>
<td>0</td>
</tr>
<tr>
<td>$t_{\text{final}}$</td>
<td>50</td>
</tr>
<tr>
<td>$\Delta t$</td>
<td>$10^{-3}$</td>
</tr>
<tr>
<td>$X_{\text{min}}$</td>
<td>0</td>
</tr>
<tr>
<td>$X_{\text{max}}$</td>
<td>50</td>
</tr>
<tr>
<td>$V_{\text{min}}$</td>
<td>$0.20 \cdot 10^{-16}$</td>
</tr>
<tr>
<td>$V_{\text{max}}$</td>
<td>$13.20 \cdot 10^{-16}$</td>
</tr>
<tr>
<td>$\Delta V$</td>
<td>$0.05 \cdot 10^{-16}$</td>
</tr>
<tr>
<td>$V_0$</td>
<td>$5 \cdot 10^{-16}$</td>
</tr>
<tr>
<td>$X_0$</td>
<td>5</td>
</tr>
</tbody>
</table>

Parameters of the multiple MC runs

<table>
<thead>
<tr>
<th>Parameters of the single MC run</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t_{\text{initial}}$</td>
</tr>
<tr>
<td>$t_{\text{final}}$</td>
</tr>
<tr>
<td>$V_{\text{initial}}$</td>
</tr>
<tr>
<td>$X_{\text{initial}}$</td>
</tr>
<tr>
<td>$\Delta t_{\text{sampling}}$</td>
</tr>
</tbody>
</table>

invariant (stationary) behavior. Indeed, it was observed that the probability profile also remains invariant for large times. This stationary probability profile satisfies the M-Equation with the time derivative equal to zero and is shown in Figure 2.3b.

**MC Simulations**

MC simulations were performed with the algorithm described in section 2.5 (with appropriate modifications so that no duplication events are considered). The stochastic paths from these simulations were sampled appropriately and the probability in the $(X,V)$ subspace of interest was estimated. Two different MC simulation schemes were used: (i) multiple simulations with same beginning and ending times and different random seeds; in this case sampling was performed at the end of the simulation. (ii) one long transient simulation, where samples were taken in constant time intervals.
Figure 2.3: Panel (a): Transient behavior of the mean molecular content and the mean volume for the first 25 min (both computed by the transient solution of the M-Equation). Panel (b): Stationary probability distribution (t = 50 min) computed by solving the M-Equation. Panel (c): Representative stochastic paths for the molecular content and the volume for the first 25 min (both computed by MC simulation). Panel (d): Stationary probability distribution (t = 50 min) computed by multiple MC simulations and sampling at the end of the simulation.

For the calculation of the probability profile from the MC simulations, we need to define sampling bins in the two dimensional state space. To this end, we utilized a mesh of bins identical to that used in the M-Equation solver. Thus, the bin \([k, m]\) corresponds to \(X = k - 1\) and \(V = (m - 1)\cdot\Delta V\) (the bins are numbered beginning with 1). Sample counting was done as follows: the sample \((X, V)\) was counted towards bin \([\min(X + 1, i_{\text{max}} + 1), \min(\fix((V - V_{\text{min}})/\Delta V) + 1, j_{\text{max}} + 1)\]).

Representative transient simulations obtained from the second simulation scheme (one long MS simulation) are shown in Figure 2.3c. Furthermore, Figure 2.3d shows the...
estimated probability profile obtained by the first simulation scheme (multiple MC simulations). As is apparent, this probability profile agrees with that obtained previously (panel a) by solving the M-Equation.

Error Analysis

The model problem introduced in this section can also be used to test the validity of our code and simulation schemes by performing error analysis. The latter procedure can quantitatively show the agreement between the probability profiles obtained by MC simulations and that obtained from the numerical solution of the M-Equation. The numerical error was calculated as follows:

$$\text{Error}_N = \sqrt{\frac{1}{N} \sum_{X=1}^{N} \sum_{i=1}^{i_{\text{max}}} \left( p_{X,i}^{\text{MC},N} - p_{X,i}^{\text{M},\text{eq}} \right)^2 \cdot \Delta V}$$  \hspace{1cm} (2.71)$$

where $p_{X,i}^{\text{MC},N}$ is the stationary probability at bin $[X, i]$ calculated from a collection of $N$ samples obtained from the MC simulations and $p_{X,i}^{\text{M},\text{eq}}$ is the stationary probability at the same bin calculated by the M-Equation solver. For completely uncorrelated samples the error is expected to follow the relation:

$$\text{Error}_N = \frac{1}{\sqrt{N \cdot \Delta V}}$$  \hspace{1cm} (2.72)$$

The results of the convergence studies are shown in Figure 2.4. Panel (a) pertains to multiple MC runs with sampling in the final time and panel (b) pertains to one MC run and sampling over time. Clearly the convergence rate is $1/2$ as expected. Note that the uncorrelated samples obtained from the multiple MC runs (a) follow equation (2.72), whereas the samples obtained from a single run (b) show somewhat higher errors ($C/\sqrt{N \cdot \Delta V}$ with $C > 1$) due to the fact that they are correlated. This explains the
Figure 2.4: Convergence in the mean square sense. Panel (a): multiple MC runs with sampling in the final time ($t = 50$ min). Panel (b): single long MC simulation with sampling over time. In both panels the open circles denote the error calculated from simulations (using equation 2.71) and the line shows the theoretically expected behavior (equation 2.72).

difference between the error points and the line corresponding to equation (2.72), that is observed in panel (b).

In fact, the smaller the sampling timestep $\Delta t_{sp}$, the larger the correlations between subsequent samples. Thus, for small $\Delta t_{sp}$, even though we collect many samples, the information we obtain is redundant, and the error remains high. However, for larger $\Delta t_{sp}$ (much larger than the autocorrelation time of the process), subsequent samples are practically uncorrelated but we collect samples less frequently. Thus, the errors in the obtained distributions will be primarily due to the small sample size. These observations have to be taken into consideration when choosing the sampling timestep $\Delta t_{sp}$.

2.7 Conclusion

In this chapter we introduced the single cell M-Equation that takes into account cell growth, DNA duplication, and division. After deriving the M-Equation we showed the expressions for the propensity functions, the growth function and the division mechanism. Subsequently, we developed MC algorithms that can simulate stochastic
paths from this M-Equation. Finally we solved the M-Equation and performed MC simulations for a model problem. We compared the results of these simulations, thereby showing convergence and validating the MC simulation algorithms.
Chapter 3:

Extending the Modeling Approach to the Cell Population Level

In this chapter, we extend the modeling framework that developed in Chapter 2 for a single cell, to cell populations.

3.1 Definitions

In Chapter 2 we defined a vector $z$ that denotes the state of a single cell. This vector contains the numbers of molecules for the $n$ species into consideration and the volume, which is the single morphometric characteristic that we take into account. Now, if we wish to extend the framework to cell populations, we need to define a vector for the state of the overall cell population. This vector will contain one entry for the number $v$ of individuals in the population and $v \cdot \text{size}(z) = v \cdot (N+1)$ entries that represent the states of each and every cell in that population (refer to section 2.1 Preliminary Definitions). Therefore if we denote the population state vector with $w$:

$$w \in \mathbb{N}_0 \times \mathcal{Z}^{N_0}$$

We will now develop a M-Equation for the temporal evolution of probability in an ensemble, which here will be a set of cell populations. For example, if all offspring cells
have originated from a single cell then we essentially follow the dynamics of the *lineage* of that cell which is the tree structure of all the daughters originating from a mother cell (see Figure 2.1). Thus, the elements of our ensemble are essentially all the stochastically possible lineages of the initial mother cell.

It is natural to consider the probability that we randomly sample the ensemble at time $t$ picking a cell population that has $v$ individuals with individual states $z_i$, $i = 1, \ldots, v$. We denote this probability $J_v(z_1, \ldots, z_v; t)$. We use the letter $J$ in accordance to the Janossy density used in the continuous population balances (Ramkrishna 2000, page 278). It is important to note that this density is symmetric since the cells cannot be distinguished in any way other than their state. Thus, the value of $J_v(z_1, \ldots, z_v, \ldots, z_v)$ remains unaltered by permutations of the $z$s and the normalization condition for $J_v$ will be (see Appendix II for the derivation):

$$
\sum_{v \geq 0} \left\{ \sum_{x_1} \cdots \sum_{x_v} \frac{1}{v!} \cdot J_v((x_1, v_1), \ldots, (x_v, v_v); t) \right\} \, dv_1 \cdots dv_v = 1 \quad (3.2)
$$

The probability that the cell population will be extinct at time $t$ is $J_0(t)$. The probability that at time $t$ the population will have $v$ cells is given as:

$$
P[N_{\text{population}} = v] = \sum_{x_1} \cdots \sum_{x_v} \frac{1}{v!} \cdot J_v((x_1, v_1), \ldots, (x_v, v_v); t) \, dv_1 \cdots dv_v \quad (3.3)
$$

### 3.2 The Cell Population M-Equation

With the above observations in mind we are ready to write the M-Equation that will describe the evolution of the probability distribution for a population of cells. In order to correctly derive each term we need to keep in mind that $J_v((x_1, v_1), \ldots, (x_i, v_i), \ldots, (x_v, v_v); t)$ behaves as probability mass function in the species
content coordinates but as probability density in the volume coordinates. Thus, the M-
Equation, which is essentially a probability balance, will be written as:

\[
J_v((X_1, V_1), ..., (X_i, V_i), ..., (X_v, V_v); t + \Delta t) \cdot \prod_{k=1}^{v} \Delta V_k
- J_v((X_1, V_1), ..., (X_i, V_i), ..., (X_v, V_v); t) \cdot \prod_{k=1}^{v} \Delta V_k = \sum_{k=1}^{4} \text{ProbIn}_k - \sum_{k=1}^{4} \text{ProbOut}_k \tag{3.4}
\]

Let us derive term by term the probability inflows and outflows. In this process we will only consider single events occurring at the time interval \([t, t+\Delta t]\), since the probability of two events happening in this interval is \(O(\Delta t^2)\).

3.2.1 Probability Flow due to reaction

We have denoted with \(a_j(X,V)\) the propensity of reaction \(j\) and with \(v_j\) the change in molecular content that reaction \(j\) brings about (section 2.2.1 Chemical Reactions). Now, the inflow of probability to state \((v, (X_1, V_1), ..., (X_i, V_i), ..., (X_v, V_v))\) will be a sum of the contributions of each cell \(\zeta\) that exists in state \((X_\zeta - v_j, V_\zeta)\) and undergoes one reaction event of type \(j\) in the next \(\Delta t\). Therefore:

\[
\text{ProbIn}_j = \sum_{\zeta=1}^{v} \sum_{j=1}^{m} a_j(X_\zeta - v_j, V_\zeta) \cdot \Delta t \cdot J_v((X_1, V_1), ..., (X_i, V_i), ..., (X_v, V_v); t) \cdot \prod_{k=1}^{v} \Delta V_k \tag{3.5}
\]

The outflow of probability due to reaction contains contributions from the cells that exist in state \((v, (X_1, V_1), ..., (X_i, V_i), ..., (X_v, V_v))\) and undergo any reaction event:

\[
\text{ProbOut}_j = \sum_{\zeta=1}^{v} \sum_{j=1}^{m} a_j(X_\zeta, V_\zeta) \cdot \Delta t \cdot J_v((X_1, V_1), ..., (X_i, V_i), ..., (X_v, V_v); t) \cdot \prod_{k=1}^{v} \Delta V_k \tag{3.6}
\]
3.2.2 Probability Flow due to DNA duplication

We have denoted with $a_s(X, V)$ the propensity of DNA duplication (subscript “s” for “synthesis”) and with $v_s$ the change in molecular content that duplication brings about (section 2.2.2 DNA Duplication). Similarly to the case of stochastic reaction occurrence, the probability inflow to state $\left( v_s, (X_1, V_1), ..., (X_i, V_i), ..., (X_v, V_v) \right)$ will be a sum of the contributions of each cell $\zeta$ that exists in state $\left( X_{\zeta}, v_s, V_{\zeta} \right)$ and undergoes one duplication event in the next $\Delta t$. Therefore:

$$\text{ProbInfl}_2 = \sum_{\zeta=1}^{v} a_s (X_{\zeta} - v_s, V_{\zeta}) \cdot \Delta t \cdot J_v (\left( X_1, V_1 \right), ..., \left( X_i - v_s, V_i \right), ..., \left( X_v, V_v \right); t) \cdot \prod_{k=1}^{v} \Delta V_k \quad (3.7)$$

The outflow of probability due to duplication contains contributions from the cells that exist in state $\left( v_s, (X_1, V_1), ..., (X_i, V_i), ..., (X_v, V_v) \right)$ and undergo DNA duplication:

$$\text{ProbOutfl}_2 = \sum_{\zeta=1}^{v} a_s (X_{\zeta}, V_{\zeta}) \cdot \Delta t \cdot J_v (\left( X_1, V_1 \right), ..., \left( X_i, V_i \right), ..., \left( X_v, V_v \right); t) \cdot \prod_{k=1}^{v} \Delta V_k \quad (3.8)$$

3.2.3 Probability flow due to growth

We have denoted with $g(X, V)$ the growth rate of the cell (section 2.2.3 Growth). Then $g(X_{\zeta}, V_{\zeta}) \cdot \Delta t$ is the volume change that will occur in cell $\zeta$ during the infinitesimal time interval $(t, t + \Delta t)$. Therefore the inflow of probability due to growth for any cell is:

$$\text{ProbIn}_3 = \sum_{\zeta=1}^{v} g (X_{\zeta}, V_{\zeta}) \cdot \Delta t \cdot J_v (\left( X_1, V_1 \right), ..., \left( X_\zeta, V_\zeta \right), ..., \left( X_v, V_v \right); t) \cdot \prod_{k=1}^{v} \Delta V_k \quad (3.9)$$

The outflow of probability due to growth will be due to the growth of cells existing in states $(X_{\zeta}, V_\zeta + \Delta V_\zeta)$:
\[ \text{ProbOut}_3 = \sum_{\zeta=1}^{3} g\left(X_{\zeta}, V_{\zeta} + \Delta V_{\zeta}\right) \Delta t \cdot J_\nu \left( \left(X_{1}, V_{1}, \ldots, X_{\nu}, V_{\nu} + \Delta V_{\nu}\right) \right) \prod_{k=1, k\neq \zeta}^{\nu} \Delta V_k \quad (3.10) \]

3.2.4 Inflow due to division

We have denoted with \( a_d(X,V) \) the propensity of division (section 2.2.4 Division). Whenever a cell divides, the number of cells in the population increases by one. Thus, the probability inflow to state \( \left(v, (X_1, V_1), \ldots, (X_\nu, V_\nu) \right) \) caused by division, will come from cell populations that exist in some state \( \left(v - 1, (Y_1, U_1), \ldots, (Y_{v-1}, U_{v-1}) \right) \) at time \( t \). Now, since one cell divides to two daughters, this means that two of the cells that exist in \( \left(v, (X_1, V_1), \ldots, (X_\nu, V_\nu) \right) \) at time \( t + \Delta t \), were a single cell within \( \left(v - 1, (Y_1, U_1), \ldots, (Y_{v-1}, U_{v-1}) \right) \) at time \( t \). Furthermore, upon division a mother cell with state \( (X_\zeta + X_\zeta, V_\zeta + V_\zeta) \) can produce a daughter cell of state \( (X_\zeta, V_\zeta) \) with probability \( h\left(X_\zeta, V_\zeta \left| X_\zeta + X_\zeta, V_\zeta + V_\zeta \right. \right) \).

In order to understand exactly what this probability represents let us consider an ensemble of dividing cells existing in state \( (X_\zeta + X_\zeta, V_\zeta + V_\zeta) \). Upon division, we arbitrarily number the daughters and look at daughter 1. Then, the fraction of them that will produce daughter 1 at state \( (X_\zeta, V_\zeta) \) is given as \( h\left(X_\zeta, V_\zeta \left| X_\zeta + X_\zeta, V_\zeta + V_\zeta \right. \right) \).

Apparently \( h \) is a symmetric function since whenever we get daughter 1 being in state \( (X_\zeta, V_\zeta) \) we necessarily get daughter 2 being in state \( (X_\zeta, V_\zeta) \). Since we number the daughter cells arbitrarily, the event that we get daughter 1 in state \( (X_\zeta, V_\zeta) \) is equally
probable to the event that we get daughter 1 in state \((X_t, V_c)\). This symmetry is
expressed as (see also equation 2.17):

\[
h(X_t, V_c | X_t + X_c, V_c + V_c) = h(X_{t-1}, V_{c-1} | X_t + X_c, V_c + V_c)
\]  
(3.11)

Now, consider an ensemble of cell populations that contain \(v - 1\) cells, exactly one of
which exists in state \((X_t + X_c, V_c + V_c)\) and is dividing in the next \(\Delta t\) time units. The
fraction of populations that will result in state \((v, (X_1, V_1), ..., (X_v, V_v))\) is equal to the
fraction of those in which the diving cell will produce daughter 1 in state \((X_t, V_c)\) plus
the fraction of those with daughter 1 in state \((X_t, V_c)\). This value can also be thought of
as the probability of the event that daughter 1 will be in state \((X_t, V_c)\) or in state
\((X_t, V_c)\). Since the two latter events are disjoint, their probabilities are summed. Thus:

\[
\text{Prob}_{in} = \sum_{\zeta = 1}^{v-1} \sum_{\zeta' = 1}^{v} a_d(X_t + X_c, V_c + V_c) \cdot \Delta t
\]

\[
\cdot \left[ h(X_t, V_c | X_t + X_c, V_c + V_c) \cdot \Delta V_c \cdot \prod_{k=1}^{v} \Delta V_k
\right]
\]

\[
\cdot \left[ -J_{\zeta-1} (X_1, V_1), ..., (X_{\zeta-1}, V_{\zeta-1}), (X_{\zeta+1}, V_{\zeta+1}), ..., (X_v, V_v); t
\right]
\]

\[
+ h(X_t, V_c | X_t + X_c, V_c + V_c) \cdot \Delta V_c \cdot \prod_{k=1}^{v} \Delta V_k
\]

\[
\cdot \left[ -J_{v-1} (X_1, V_1), ..., (X_{\zeta-1}, V_{\zeta-1}), (X_{\zeta+1}, V_{\zeta+1}), ..., (X_v, V_v); t
\right]
\]

(3.12)

and due to the already discussed symmetry properties of \(h\) and \(J\) we can simplify the
above expression as follows:
\[ \text{ProbIn}_4 = 2 \cdot \sum_{\zeta=1}^{\nu-1} \sum_{\zeta_{\nu-1}}^\nu a_d \left( X_\zeta + X_{\zeta}, V_\zeta + V_{\zeta} \right) \cdot \Delta t \]

\[ \cdot h \left( X_\zeta, V_\zeta \big| X_\zeta + X_{\zeta}, V_\zeta + V_{\zeta} \right) \cdot \prod_{k=1}^\nu \Delta V_k \]

\[ \cdot J_{\nu-1} \left( \left( X_1, V_1 \right), \ldots, \left( X_\zeta + X_{\zeta}, V_\zeta + V_{\zeta} \right), \ldots, \left( X_{\zeta-1}, V_{\zeta-1} \right), \left( X_{\zeta+1}, V_{\zeta+1} \right), \ldots, \left( X_\nu, V_\nu \right); t \right) \]

(3.13)

The outflow of probability due to division is:

\[ \text{ProbOut}_4 = \sum_{\zeta=1}^\nu a_d \left( X_\zeta, V_\zeta \right) \cdot \Delta t \cdot J_\nu \left( \left( X_1, V_1 \right), \ldots, \left( X_\zeta, V_\zeta \right), \ldots, \left( X_\nu, V_\nu \right); t \right) \cdot \prod_{k=1}^\nu \Delta V_k \]  

(3.14)

3.2.5 Overall M-Equation

From equation (3.4), by substituting the probability inflow and outflow terms from expressions (3.5, 6), (3.7, 8), (3.9, 10), and (3.13, 14), dividing with \( \prod_{k=1}^\nu \Delta V_k \cdot \Delta t \) and taking limits as \( \Delta V_k \to 0 \), \( \Delta t \to 0 \) we end up with the following M-Equation:

\[ \frac{\partial}{\partial t} J_\nu \left( \left( X_1, V_1 \right), \ldots, \left( X_\zeta, V_\zeta \right), \ldots, \left( X_\nu, V_\nu \right); t \right) = \]

\[ \sum_{\zeta=1}^\nu \sum_{j=1}^m \left[ a_j \left( X_\zeta - v_j, V_\zeta \right) \cdot J_\nu \left( \left( X_1, V_1 \right), \ldots, \left( X_\zeta - v_j, V_\zeta \right), \ldots, \left( X_\nu, V_\nu \right); t \right) - a_j \left( X_\zeta, V_\zeta \right) \cdot J_\nu \left( \left( X_1, V_1 \right), \ldots, \left( X_\zeta, V_\zeta \right), \ldots, \left( X_\nu, V_\nu \right); t \right) \right] \]

\[ - \sum_{\zeta=1}^\nu \sum_{j=1}^m \left[ a_j \left( X_{\zeta-1} - v_j, V_\zeta \right) \cdot J_\nu \left( \left( X_1, V_1 \right), \ldots, \left( X_{\zeta-1} - v_j, V_\zeta \right), \ldots, \left( X_\nu, V_\nu \right); t \right) \right.

\[ - a_j \left( X_\zeta, V_\zeta \right) \cdot J_\nu \left( \left( X_1, V_1 \right), \ldots, \left( X_{\zeta-1}, V_{\zeta-1} \right), \ldots, \left( X_\nu, V_\nu \right); t \right) \]

\[ - \sum_{\zeta=1}^\nu \frac{\partial}{\partial V_\zeta} \left[ g \left( X_\zeta, V_\zeta \right) \cdot J_\nu \left( \left( X_1, V_1 \right), \ldots, \left( X_\zeta, V_\zeta \right), \ldots, \left( X_\nu, V_\nu \right); t \right) \right] \]

\[ + 2 \cdot \sum_{\zeta=1}^\nu \sum_{\zeta_{\nu-1}}^\nu a_d \left( X_\zeta + X_{\zeta}, V_\zeta + V_{\zeta} \right) \cdot h \left( X_\zeta, V_\zeta \big| X_\zeta + X_{\zeta}, V_\zeta + V_{\zeta} \right) \]

\[ \cdot J_{\nu-1} \left( \left( X_1, V_1 \right), \ldots, \left( X_\zeta + X_{\zeta}, V_\zeta + V_{\zeta} \right), \ldots, \left( X_{\zeta-1}, V_{\zeta-1} \right), \left( X_{\zeta+1}, V_{\zeta+1} \right), \ldots, \left( X_\nu, V_\nu \right); t \right) \]

\[ - \sum_{\zeta=1}^\nu a_d \left( X_\zeta, V_\zeta \right) \cdot J_\nu \left( \left( X_1, V_1 \right), \ldots, \left( X_\zeta, V_\zeta \right), \ldots, \left( X_\nu, V_\nu \right); t \right) \]  

(3.15)
3.3 MC Simulation of the Cell Population M-Equation

Consider a population that at time $t$ has $v$ cells, each at state $(X_i, V_i)$ for $i = 1, \ldots, v$. The cells are assumed not to interact with each other, and thus we can compute a division and a reaction waiting time for each one of the cells. For each cell $i$ ($i = 1, \ldots, v$), we can compute the waiting times for the occurrence of a reaction, a DNA duplication, or a division event. The event that will be simulated first will be the one with the shortest waiting time $\tau$. If the event was a division, then both of the newborn cells will be taken into account by increasing the cell population number by one and storing the state of both newborn cells in the population state vector (needless to say, one daughter will replace the mother cell). If the event was a reaction or a DNA duplication, the state of the individual cell in which the reaction happened will be updated. Thus, the new state of the population will reflect the reaction duplication or division event and the waiting times for the occurrences of the reaction or division events will be updated.

A few fundamental observations can be made that can be used to optimize the algorithm:

1. The waiting times of one cell are independent of those of all other cells. Thus, suppose that we have a population of $v$ cells at time $t$, and the algorithm determines that a reaction or duplication event will happen in cell $i$ at time $t + \tau$. Then, the state of cell $i$ will change, and we will have to calculate new division and reaction waiting times for cell $i$. However, for all the other cells, we can merely subtract $\tau$ from their waiting times, and the resulting waiting times will follow the correct probability distributions for the respective event occurrences. We can use this strategy because there is no interaction between the cells in the population, and thus if a reaction
happens to cell i, the other cells are not affected. This assumption would not hold if
the cells interacted with each other by an extracellular messenger and the event to be
simulated was secretion of one messenger molecule to the extracellular space; in this
case, all cells are affected by this event. Likewise, if a division event happens, we will
have to generate waiting times for the reaction and division events of the two
daughter cells; for the other cells, we can merely subtract \( \tau \) from their waiting times.
Furthermore, we can eliminate the need for subtracting \( \tau \) each time an event occurs
by working on absolute rather than relative time (thus, we always store the absolute
time for the occurrence of the events of interest).

2. Since a duplication event always precedes a division event, we can set the duplication
time to infinity immediately after a duplication event and recalculate it after a
division event. Likewise, we can set the division times of the two daughters to infinity
immediately after their birth and recalculate them after a duplication event.

3. Once we have taken care of the precedence of the duplication to division, the waiting
times for division and duplication are only volume dependent, and volume evolves
deterministically. Thus, we can calculate duplication or division waiting times only
once (after a division or a duplication event, respectively). We do not have to update
them every time a reaction event happens. This case would not hold, however, if the
division or duplication propensities were also dependent on the copy numbers of
other species in the cell.

4. The above observations enable us to make the minimum possible updates to the
vectors containing the absolute times for the occurrences of the events to be
simulated. Evidently, after each simulated event, we will have to update at most
2-(m+2) absolute times (reaction, division and duplication times for two newborn cells) and to find the minimum between all times. Both tasks can be efficiently accomplished by using heap structures (binary trees). The absolute times for reaction, division and duplication are stored in three different heaps. Sorting in the heap occurs automatically upon update or addition of a new time after simulation of reaction/duplication or division events respectively.

3.3.1 Pseudo-Code

We assume that the following are known: the reaction network and all reaction propensities, the duplication propensity, the growth function g, the division propensity and the partitioning mechanism. Then the outline of the procedures to be simulated is as follows. Comments are denoted as :comment:.

1. Begin execution

:initializations begin:

2. Define the initial and final times of the simulation. Initialize time as t = initial time

3. Define the initial number of cells ν in the population and the initial molecular contents Xᵢ and volume Vᵢ of the cells i = 1,...,ν

4. Define a constant Δtₛₚ for sampling through time, and define the time that the first sample will be taken tₛₚ = initial time

5. Initialize a filename counter: cntr = 0

6. Do for every cell in the population i = 1,...,ν

6.1. Calculate the waiting time for the next reaction event, τᵢ, and store the absolute time tᵢ = t + τᵢ in the heap for the reaction times
6.2. Calculate the waiting time for the next duplication event, $\tau_{\text{dup},i}$ and store the absolute time $t_{\text{dup},i} = t + \tau_{\text{dup},i}$ in the heap for the duplication times

6.3. Calculate the waiting time for the next division event, $\tau_{\text{div},i}$ and store the absolute time $t_{\text{div},i} = t + \tau_{\text{div},i}$ in the heap for the division times

7. Store the old population size $v_{\text{old}} = v$

8. Initialize the indicator of the affected cell (namely the cell in which the next event will be simulated) $v_{\text{cell}} = v$

:initializations end:

:main simulation block begins:

9. Do while $t \leq$ final time

:updating of the previous state of the cell population begins:

9.1. If $v_{\text{old}} = v$, then update the species copy numbers of the affected cell $v_{\text{cell}}$:

\[ X_{\text{old,cell}} = X_{v_{\text{cell}}} \]

9.2. If $v_{\text{old}} < v$, then update the species copy numbers of the affected cell $v_{\text{cell}}$ and the newborn cell $v$: $X_{\text{old,cell}} = X_{v_{\text{cell}}}, X_{\text{old,v}} = X_v$

9.3. Update previous volumes of all cells $V_{\text{old,i}} = V_i, i = 1, \ldots, v$

9.4. Store previous time and population $t_{\text{old}} = t, v_{\text{old}} = v$

:updating of the previous state of the cell population ends:

:determination of the next simulated event begins:

9.5. Find the minimum absolute time for the next reaction event, $t_{\text{rxn,rxn}}$ and the corresponding cell $j_{\text{rxn}}$

9.6. Find the minimum absolute time for the next DNA duplication event, $t_{\text{dup,dup}}$ and the corresponding cell $j_{\text{dup}}$
9.7. Find the minimum absolute time for the next division event, $t_{\text{div,jdiv}}$ and the corresponding cell $j_{\text{div}}$

9.8. If $t_{\text{rxn,jrxn}} < t_{\text{dup,jdup}}$ and $t_{\text{rxn,jrxn}} < t_{\text{div,jdiv}}$, then

9.8.1. Reaction event is going to happen; set $\nu_{\text{cell}} = j_{\text{rxn}}$

9.9. Else if $t_{\text{dup,jdup}} < t_{\text{rxn,jrxn}}$ and $t_{\text{dup,jdup}} < t_{\text{div,jdiv}}$, then

9.9.1. DNA duplication event is going to happen; set $\nu_{\text{cell}} = j_{\text{dup}}$

9.10. Else

9.10.1. Division event is going to happen; set $\nu_{\text{cell}} = j_{\text{div}}$

\textit{determination of the next simulated event ends:}

\textit{simulation of the next event begins:}

9.11. Select case for the next simulated event

9.11.1. Reaction event:

9.11.1.1. Find $\mu$ the kind of reaction to be simulated

9.11.1.2. Update volumes of all cells $V_i$ according to the growth law

9.11.1.3. Simulate the reaction event, update the state $X_{j_{\text{rxn}}}$

9.11.1.4. Update absolute time $t = t_{\text{rxn,\nu_{cell}}}$

9.11.1.5. Update the old absolute reaction time for the affected cell $\nu_{\text{cell}}$ with a newly generated one in the corresponding heap

9.11.2. Duplication event:

9.11.2.1. Update volumes of all cells $V_i$ according to the growth law

9.11.2.2. Update absolute time $t = t_{\text{dup,\nu_{cell}}}$

9.11.2.3. Update the affected cell contents according to the duplication event
9.11.2.4. Update the old absolute reaction time for the affected cell \( v_{cell} \) with a newly generated one in the corresponding heap

9.11.2.5. Update the old (infinite) division time for the affected cell \( v_{cell} \) with a newly generated one in the corresponding heap

9.11.2.6. Set duplication time to infinity for the affected cell \( v_{cell} \)

9.11.3. Division event:

9.11.3.1. Update volumes of all cells \( V_i \) according to the growth law, up to the point just before division \( t_{div,v_{cell}} \)

9.11.3.2. Update absolute time \( t = t_{div,v_{cell}} \)

9.11.3.3. Update the cell population \( v = v + 1 \)

9.11.3.4. Find \( \rho \) the ratio of the volumes of the first daughter to the mother cell

9.11.3.5. Find \( \xi_i \) the number of molecules of species \( i \) that the first daughter cell \( v_{cell} \) will inherit

9.11.3.6. Calculate the states of the first daughter \( X_{v_{cell}}, V_{v_{cell}} \) and the second daughter \( X_v, V_v \)

9.11.3.7. For the affected cell \( v_{cell} \): update the old absolute reaction and duplication times with newly generated ones in the corresponding heaps, and set division time to infinity

9.11.3.8. For the new daughter cell \( v \): insert newly generated absolute reaction, duplication times into the corresponding heaps, and an infinite division time into the division-time-heap.

:simulation of the next event ends:
writing the state of the population in $\Delta t_{sp}$ time intervals begins:

9.12. Do while $t_{sp} < t$:

9.12.1. Open a file $\text{filename}_{cntr}$ for writing population data

9.12.2. write into $\text{filename}_{cntr}$ data $t_{sp}$, $v$

9.12.3. Do for every cell in the population $i = 1, \ldots, v$

9.12.3.1. write into $\text{filename}_{cntr}$ data $X_{old,i}$, $V_{i}(t_{sp})$

9.12.4. Close the file $\text{filename}_{cntr}$

9.12.5. Update sampling time: $t_{sp} = t_{sp} + \Delta t_{sp}$

9.12.6. Update filename counter: $cntr = cntr + 1$

writing the state of the population in $\Delta t_{sp}$ time intervals ends:

main simulation block ends:

10. Terminate execution

Comments:

1. In step 9.12.3.1 one has to be careful to write $V$ corresponding to the time $t_{sp}$. Thus, if the volume at time $t + \tau$ is given as a function of the volume $V(t)$, the state $X$ at time $t$ and the waiting time $\tau$, $\Phi(X, V(t), \tau)$, then in the file one should write the following value for the volume: $V_{i}(t_{sp}) = \Phi(X_{old}, V_{old}, t_{sp} - t_{old})$.

2. The use of heap structures is facilitated by having separate subroutines that handle the heap structure and perform the insert or update operations.

3.4 A Simple Model-Problem

In the previous sections we extended the modeling approach of Chapter 2 to cell populations. The resulting M-Equation described the evolution of the probability of
finding at time $t$ a cell population consisting of $v$ individuals which have the states $(X_1, V_1), (X_2, V_2), \ldots, (X_v, V_v)$. We also developed a MC algorithm for the simulation of stochastic paths of this M-Equation.

In this section we will demonstrate this algorithm by simulating a genetic network that consists of two genes under the influence of two identical repressible promoters. Such a genetic network was used by Elowitz et al. (2002) to decompose the extrinsic and intrinsic contributions of noise to the overall single cell noise. In particular, two GFP variants, a yellow (YFP) and a cyan (CFP), were cloned in opposite positions from the origin of replication into the *E. coli* chromosome. Expression of both proteins is driven from identical Lac repressible promoters, and the fluorescence intensity of both variants is approximately the same.

Thus, measurements of the fluorescence of the cells in the two different channels, yellow and cyan, can give indications of the intrinsic and the extrinsic noise. In particular, difference in the fluorescence of the two channels for the same cell originates from the intrinsic noise, and difference in the fluorescence between distinct cells is a result of the extrinsic noise. This decomposition of noise to extrinsic and intrinsic components is rather phenomenological and based on how the protein contents of identical genes correlate. However, using the framework developed in sections 3.1 - 3.3 we can identify the contribution of fundamental biological mechanisms on the extrinsic or the intrinsic noise components.

3.4.1 Reaction Network; Growth, Duplication and Division Mechanisms

In order to model this system, we consider a set of reactions in which the following species participate, RNA polymerase (RP), ribosome (RB), Lac repressor (Lac), free yfp
operator ($O_{yfp}$), Lac repressed $yfp$ operator ($O_{yfp,Lac}$), $yfp$ mRNA ($R_{yfp}$), Yfp protein molecule (Yfp), free $cfp$ operator ($O_{cfp}$), Lac repressed $cfp$ operator ($O_{cfp,Lac}$), $cfp$ mRNA ($R_{cfp}$), Cfp protein molecule (Cfp). Note that we have two chromosomal DNA species (the two operators) each of which can exist in two states namely $O_{yfp}$, $O_{yfp,Lac}$ and $O_{cfp}$, $O_{cfp,Lac}$. Therefore the species vector is:

$$S = \{RP, RB, Lac, R_{yfp}, Yfp, R_{cfp}, Cfp, O_{yfp}, O_{yfp,Lac}, O_{cfp}, O_{cfp,Lac}\}$$  \hspace{1cm} (3.16)

Assuming 0th order kinetics for the production of RNA polymerase (RP), ribosome (RB) and Lac repressor molecules, the corresponding reactions are:

$$\emptyset \xrightarrow{k_1} RP$$  \hspace{1cm} (3.17)

$$\emptyset \xrightarrow{k_2} RB$$  \hspace{1cm} (3.18)

$$\emptyset \xrightarrow{k_3} Lac$$  \hspace{1cm} (3.19)

For the repressive action of the Lac to the $yfp$ operator we write the following chemical equilibrium:

$$O_{yfp} + Lac \xrightleftharpoons[k_4]{k_3} O_{yfp,Lac}$$  \hspace{1cm} (3.20, 21)

The transcription and translation of $yfp$ are assumed to follow 2nd order kinetics:

$$O_{yfp} + RP \xrightarrow{k_5} O_{yfp} + RP + R_{yfp}$$  \hspace{1cm} (3.22)

$$R_{yfp} + RB \xrightarrow{k_6} R_{yfp} + RB + YFP$$  \hspace{1cm} (3.23)

Similarly for $cfp$:

$$O_{cfp} + Lac \xrightleftharpoons[k_7]{k_6} O_{cfp,Lac}$$  \hspace{1cm} (3.24, 25)

$$O_{cfp} + RP \xrightarrow{k_7} O_{cfp} + RP + R_{cfp}$$  \hspace{1cm} (3.26)
\[ R_{cfp} + RB \xrightarrow{k_2} R_{cfp} + RB + CFP \]  

(3.27)

Degradation reactions are assumed to follow first order kinetics:

\[ \text{RP} \xrightarrow{k_I} \emptyset \]  

(3.28)

\[ \text{RB} \xrightarrow{k_{II}} \emptyset \]  

(3.29)

\[ \text{Lac} \xrightarrow{k_{III}} \emptyset \]  

(3.30)

\[ \text{R}_{yfp} \xrightarrow{k_{IV}} \emptyset \]  

(3.31)

\[ \text{YFP} \xrightarrow{k_{V}} \emptyset \]  

(3.32)

\[ \text{R}_{cfp} \xrightarrow{k_{VI}} \emptyset \]  

(3.33)

\[ \text{CFP} \xrightarrow{k_{VII}} \emptyset \]  

(3.34)

The expressions of the reaction propensity functions are constructed using general formula (2.23) according to the reaction scheme above. For the simulation of the single cell growth process, expression (2.27) is used (exponential growth). The duplication propensity is taken as in expression (2.29). At every division event the total operator contents for \( yfp \) and \( cfp \) are doubled by introducing free operator contents equal to \( O_{yfp, \text{Total}} \) and \( O_{cfp, \text{Total}} \). The division propensity is given by equation (2.30) and the partitioning mechanism is given by equations (2.31, 32, 36, 38).

3.4.2 Simulation Results

The system was simulated using several different parameter sets in order to elucidate the effect of each mechanism on the overall, as well as extrinsic and intrinsic noise (heterogeneity) in the cell population. Thus, the parameter values for the nominal set appear in Table 3.1 and when different parameter values are used, it is so noted in the particular figure caption. Note that for the nominal parameter set, no repression is
Table 3.1: Parameter values for the two promoter system (no repressive action)

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Units</th>
<th>Value</th>
<th>Symbol</th>
<th>Units</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$k_1$</td>
<td>(nM·min$^{-1}$)</td>
<td>480</td>
<td>$k_{12}$</td>
<td>(min$^{-1}$)</td>
<td>0.01</td>
</tr>
<tr>
<td>$k_2$</td>
<td>(nM·min$^{-1}$)</td>
<td>850</td>
<td>$k_{13}$</td>
<td>(min$^{-1}$)</td>
<td>0.4</td>
</tr>
<tr>
<td>$k_3$</td>
<td>(nM·min$^{-1}$)</td>
<td>0</td>
<td>$k_{14}$</td>
<td>(min$^{-1}$)</td>
<td>0.01</td>
</tr>
<tr>
<td>$k_4$</td>
<td>(nM$^{-1}$·min$^{-1}$)</td>
<td>240</td>
<td>$k_{15}$</td>
<td>(min$^{-1}$)</td>
<td>0.4</td>
</tr>
<tr>
<td>$k_{-4}$</td>
<td>(min$^{-1}$)</td>
<td>2.4</td>
<td>$k_{16}$</td>
<td>(min$^{-1}$)</td>
<td>0.01</td>
</tr>
<tr>
<td>$k_5$</td>
<td>(nM$^{-1}$·min$^{-1}$)</td>
<td>1.2·10$^{-2}$</td>
<td>$O_{cfp_{Total}}$ (molec.)</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>$k_6$</td>
<td>(nM$^{-1}$·min$^{-1}$)</td>
<td>1.3·10$^{-7}$</td>
<td>$O_{yfp_{Total}}$ (molec.)</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>$k_7$</td>
<td>(nM$^{-1}$·min$^{-1}$)</td>
<td>240</td>
<td>$g$</td>
<td>(min$^{-1}$)</td>
<td>0.0231</td>
</tr>
<tr>
<td>$k_{-7}$</td>
<td>(min$^{-1}$)</td>
<td>2.4</td>
<td>$n_d$</td>
<td>(dim/less)</td>
<td>20</td>
</tr>
<tr>
<td>$k_8$</td>
<td>(nM$^{-1}$·min$^{-1}$)</td>
<td>1.2·10$^{-2}$</td>
<td>$V_{d,crit}$ (L)</td>
<td>$11·10^{-16}$</td>
<td></td>
</tr>
<tr>
<td>$k_9$</td>
<td>(nM$^{-1}$·min$^{-1}$)</td>
<td>1.3·10$^{-7}$</td>
<td>$q$</td>
<td>(dim/less)</td>
<td>80</td>
</tr>
<tr>
<td>$k_{10}$</td>
<td>(min$^{-1}$)</td>
<td>0.01</td>
<td>$n_s$</td>
<td>(dim/less)</td>
<td>20</td>
</tr>
<tr>
<td>$k_{11}$</td>
<td>(min$^{-1}$)</td>
<td>0.01</td>
<td>$V_{s,crit}$ (L)</td>
<td>$8·10^{-16}$</td>
<td></td>
</tr>
</tbody>
</table>

considered.

The ICs for all simulations were constructed by solving the corresponding
deterministic model and converting the concentrations to numbers of molecules. Then a
population of a single cell ($v = 1$) having those molecular contents is used as the IC.

Alternatively, one can also simulate a single cell line prior to simulating the whole cell
population. After sufficient time passes so that the process reaches time invariance, the
state of the cell is recorded. Then a population of a single cell ($v = 1$) having that
recorded state, can be used as IC for the simulation of the population. The results of the
simulations are discussed below.

**Nominal Parameter Set**

The parameter values for the nominal parameter set (Table 3.1) were chosen such that
the simulation results agree qualitatively with the results in Elowitz et al (2002). Figure
3.1 shows a simulation for this parameter set, for which all sources of noise that can be
captured with our model are present. Thus, panel (a) shows transients for the volume and
cfp content, and it is apparent that there is considerable stochasticity in both the content
Figure 3.1: Panel (a): Transient behavior of the Cfp content and the volume for a single cell line out of the cell population. Panel (b): Transient behavior of the average Cfp content of the cell population. Panel (c): Number of cells in the population as a function of time. Panel (d): Normalized Yfp content with respect to the normalized Cfp content. Each point represents one cell. Color coding corresponds to density of points. Nominal parameter set (Table 3.1).

timecourse as well as the division times. Panel (b) portrays the population average Cfp content with respect to time and panel (c) the number of individual cells in the population. For low numbers of cells, the average content oscillates following the dynamics of the division. However, as the more cells are born, their divisions occur in a much less synchronized fashion and the population average tends to a constant value.

Finally panel (d) shows the normalized Yfp content versus the normalized Cfp content in a plot similar to that used by Elowitz et al. (2002). In this plot each point corresponds to one cell of the population. Scatter of points is indicative of the existence of noise which results in heterogeneity in the cell population. In our case, transcriptional
and translational stochasticity is significant due to the low copy numbers of mRNA and protein; these are the intrinsic noise sources and contribute to the spread of points far from the diagonal \( C_f = Y_f \). Furthermore, the stochasticity in DNA duplication and division, as well as the fluctuations in the contents of RNA polymerase and ribosomes, are the extrinsic noise sources and contribute to the elongation of the ellipsoid along the diagonal \( C_f = Y_f \).

**Homogeneous Populations**

In homogeneous populations, all cells behave identically which means that (i) the fluctuations of the species copy numbers due to reactions have to be infinitesimal, (ii) duplication and division events must occur in synchrony, (iii) the cells must divide in a way that the two daughter cells have equal volumes and contents. For these three objectives to be achieved, the following conditions must be met.

First of all, the species copy numbers have to be as high as possible. For a reacting system, it is known (Schrödinger 1967) that the stochastic fluctuations in the species copy numbers are on the order of the inverse square root of the total number of interacting molecules. Thus, these fluctuations become negligible as the overall production rates of the interacting species become much larger than their respective degradation rates. Note that manipulating the copy number of the chromosomal operators has no physical significance, but in the case where no repression exists the operator is always in the unbound state, and thus it does not create any contribution to the overall noise.

Furthermore, the reactions and divisions have to be synchronized so \( n_d \) and \( n_e \) must tend to infinity in equations (2.29) and (2.30). Also, for partitioning the cells to two daughters of the same volume, parameter \( q \) in equation (2.31) must tend to infinity. Thus,
Figure 3.2: Panels (a-d) as in Figure 3.1 for the following parameter set: \( k_1 = 4800 \) nM/min, \( k_2 = 8500 \) nM/min, \( k_3 = 0.12 \) (nM-min)^{-1}, \( n_c = n_d = 10000 \). The requirement \( q \to \infty \) was numerically implemented by setting the daughter volumes equal to half the mother volume at every division. All other parameters as in Table 3.1.

the random number that expresses the ratio of volumes, \( \rho = V_d/V_m \), will take the value \( \frac{1}{2} \) almost surely. In this case, and for large copy numbers of molecules in the cell, the partitioning of the contents will then be done equally between the two daughters. Equal partitioning of the contents is guaranteed by the limiting properties of the binomial distribution: as the number of molecules to be partitioned increases to infinity, the probability in the binomial distribution (2.32) tends to accumulate to the point \( \rho \cdot X_m \) thus finally each daughter will inherit approximately half of the molecules of the mother cell.

Figure 3.2 shows a simulation of the case where no heterogeneity is observed. For this case, the single cell timecourses for Cfp content and volume appear periodic (panel a)
Figure 3.3: As in Figure 3.1d for the following parameter sets: panel (a), intrinsic noise only: $k_1 = 4800$ nM/min, $k_2 = 8500$ nM/min, $k_5 = 2 \cdot 10^{-4}$ (nM-min)$^{-1}$, $n_v = n_d = 10000$. The requirement $q \to \infty$ was numerically implemented by setting the daughter volumes equal to half the mother volume at every division; panel (b), extrinsic noise arising only from fluctuations in the RNA polymerase: $k_1 = 1.2$ nM/min, $k_2 = 8500$ nM/min, $k_5 = 480$ (nM-min)$^{-1}$, $n_v = n_d = 10000$. The requirement $q \to \infty$ was numerically implemented by setting the daughter volumes equal to half the mother volume at every division. All other parameters as in Table 3.1.

and so does the cell population average since the cells are synchronized (panel b). As a result of this synchrony, the number of cells in the population increases in a stepwise manner: in each step the number of cells is doubled (panel c). Finally, since no intrinsic or extrinsic noise is present, the Cfp and Yfp contents are identical to all cells at all times, and thus the Cfp versus Yfp plot shows that all points representing cells are concentrated to a very narrow region of the Cfp-Yfp plane.

**Only Extrinsic or Only Intrinsic Noise**

We have so far analyzed the limiting cases where the noise is negligible and where all sources of noise are significant. However, one can construct parameter sets where only extrinsic or only intrinsic noise is present. Thus, Figure 3.3a shows the Cfp versus Yfp graph in the case where only intrinsic noise is present. Stochasticity in the biomolecular reactions is significant but DNA duplication and symmetric division events occur in synchrony. In this case, the points that represent cells form a circular pattern, showing that the variability in the Cfp and Yfp content of a single cell is equal to the variability of
Cfp (or Yfp) content between different cells of the cell population.

On the other hand, Figure 3.3b pertains to a case where only extrinsic noise is present. The latter is brought about only by fluctuations in the RNA polymerase. Division is still symmetric in this case, and the duplication and division events are synchronized. Moreover, extrinsic noise is negligible because the transcriptional rates of cfp and yfp are high, thereby keeping mRNA and protein contents high. In this case, the points in the Cfp and Yfp graph are arranged along the line Cfp = Yfp, showing that in any single cell the Cfp and Yfp contents are identical, but there exists variability between different cells of the cell population.

**Effect of Repression**

In order to elucidate the effect of repression on the extrinsic and intrinsic noise, we first consider a parameter set that results in high numbers of molecules for all the species except the operators. Panel (a) of Figure 3.4 shows that for this parameter set significant extrinsic noise is observed due to the uncorrelated fluctuations of the states of the cfp and yfp operators. On the other hand, if we consider a parameter set for which all species copy numbers are high the noise becomes negligible (panel b; the total copy number for each of the two operators is 1000).

The two simulations just discussed pertain to cases where the repressor copy numbers are high and DNA duplications and cell divisions occur in synchrony; thus, no extrinsic noise was observed. However, if the repressor copy numbers are low (Figure 3.4c), then fluctuations in the state of the repressor create variability between the cells in the population (extrinsic noise) which is manifested as an ellipsoidal deformation of the cloud of points along the line Cfp = Yfp. For the simulation of panel (c) only one
operator for each protein exists. On the other hand, if the operator copy numbers are high
(panel d; \(O_{\text{yfp,Total}} = O_{\text{cfp,Total}} = 1000\)), then both intrinsic and extrinsic noise become
negligible. One might have expected that only intrinsic noise would be eliminated; yet,
since the repressor fluctuates in low copy numbers, the additional operators that it can
repress are a tiny fraction of the overall operators that exist in this case. Thus, the
fluctuations of the number of free operators are low, thereby making extrinsic as well as
intrinsic noise negligible.

---

**Figure 3.4:** As in Figure 3.1d for the following parameter sets: panel (a), \(k_1 = 4800 \text{ nM/min, } k_2 = 8500 \text{ nM/min, } k_3 = 2000 \text{ nM/min, } k_4 = 2.4 \times 10^{-5} \text{ (nM-min)}^{-1}, k_5 = 0.12 \text{ (nM-min)}^{-1}, n_a = n_d = 10000, q \rightarrow \infty\); panel (b): as in panel (a) with \(k_5 = 0.12 \times 10^{-3} \text{ (nM-min)}^{-1}, O_{\text{cfp,Total}} = O_{\text{yfp,Total}} = 1000\); panel (c): as in panel (a) with \(k_3 = 8.1 \times 10^{-2} \text{ nM/min, } k_4 = 7.2 \times 10^{-1} \text{ (nM-min)}^{-1}\); panel (d): as in panel (b) with \(k_3 = 35 \text{ nM/min, } k_5 = 0.12 \times 10^{-3} \text{ (nM-min)}^{-1}, O_{\text{cfp,Total}} = O_{\text{yfp,Total}} = 1000\). Parameters that are not mentioned have values as in Table 3.1.
3.5 Conclusion

In this chapter we extended the modeling framework was that developed in Chapter 2 for a single cell to cell populations. We first constructed the M-Equation that describes the temporal evolution of the probability of finding a cell population at time $t$ having $v$ individual cells at states $(X_1, V_1), \ldots, (X_v, V_v)$. Subsequently, we developed a MC algorithm that can be used to simulate stochastic paths of that M-Equation. Finally, we tested this algorithm in a simple model problem that refers to a genetic network with two genes under the influence of two identical repressible promoters. We demonstrated that in this case, the algorithm can successfully predict the effects of biomolecular mechanisms on the intrinsic and extrinsic noise which shapes the heterogeneity of the cell population. We also showed that our framework can be used to show the contribution of fundamental biological mechanisms on the extrinsic and intrinsic noise components. Thus, computational studies using this approach can complement experiments in which a fundamental decomposition of noise sources is impossible.
Part II: Modeling Cell Population Heterogeneity in a Genetic Network with Positive Feedback Architecture
Chapter 4:

The \textit{lac} operon: Review of Experimental and Modeling Studies

In the previous chapters we were primarily concerned about developing a generic mathematical and computational framework for modeling cell population heterogeneity. In this second part of the thesis we will apply this framework (as well as already existing tools) in modeling cell population heterogeneity in the \textit{lac} operon, a genetic network that has served as a paradigm of genetic regulation with positive feedback. We will thus start with an introduction on the previous experimental and theoretical studies of this system.

4.1 Experimental Studies

The \textit{lac} operon genetic switch is a paradigm for genetic regulation with positive feedback and has been studied experimentally and theoretically for nearly half a century. Indeed, the operon concept was first introduced in 1960 and pertains to a sequence of genes that function under the control of the same operator (Jacob et al. 1960). The \textit{lac} operon consists of three genes downstream of the \textit{lac} promoter which encode for the proteins necessary for lactose metabolism. Specifically, \textit{lacZ} encodes for the \(\beta\)-galactosidase, which transforms lactose to the inducer allolactose, as its primary function,
and digests lactose to glucose and galactose, $lacY$ encodes for the LacY permease, which transports lactose into the cell, and $lacA$ for the galactoside transacetylase (LacA), which transfers an acetyl group from acetyl-CoA to $\beta$-galactosides (Müller-Hill 1996).

Furthermore, upstream of the promoter, there exists the constitutively expressed $lacI$ gene, which encodes for the LacI repressor protein.

The free LacI repressor is a tethered tetramer (dimer of dimers) (Friedman et al. 1995) that has a high affinity for the $lacO (O1)$ operator contained in the lac promoter. Therefore, in the absence of lactose, a LacI dimer site within the tetramer binds to an operator, and thus inhibits the transcription of the $lacZ$, $lacY$ and $lacA$ genes. LacI can also bind to pseudo-operators that exist upstream (O3) and downstream (O2) of the promoter and, by virtue of its tetrameric structure, create DNA loop-structures (Oehler et al. 1990; Matthews 1992; Lewis et al. 1996). It has been suggested that the binding of the LacI repressor to the pseudo-operators, results in a localization of LacI close to the main operator thereby increasing its binding efficiency to the operator (Vilar and Saiz 2005). Thus, more efficient suppression of the lac genes is attained. Further, looped structures are more stable and ensure effective repression.

However, if lactose is present in the extracellular medium, it gets transported into the cell, where one fraction is hydrolyzed to galactose and glucose, and the other fraction is transformed to the inducer allolactose by the $\beta$-galactosidase. The allolactose binds to the LacI repressor and forms a complex with reduced binding affinity to the operator. This process results in freeing the operator site(s). Induction of LacI can also be achieved with a gratuitous inducer, such as IPTG, which does not require transformation by the $\beta$-galactosidase. Instead, it can readily bind to the repressor with a stoichiometry of two
IPTG molecules per LacI dimer (Lewis et al. 1996).

Yet, for transcription of the *lac* operon genes to be initiated, the activator cAMP-CRP complex needs to bind to a sequence near the *lac* promoter, thereby enhancing the binding affinity of the RNA polymerase (Malan et al. 1984). High activator concentrations are brought about by low glucose concentrations by altering adenylyl cyclase. Hence, the *lac* operon genes are expressed only if the glucose concentration is low and simultaneously the lactose concentration is high. Furthermore, it is known that extracellular glucose inhibits lactose transport into the cell (Postma et al. 1996; Saier et al. 1996). Therefore, glucose inhibits lactose metabolism in a dual manner: by reducing cAMP, and thus cAMP-CRP activator concentrations thereby suppressing the *lac* operon genes (catabolite repression), and by inhibiting lactose uptake (inducer exclusion).

Furthermore, since LacY facilitates lactose import resulting in repressor inactivation, it follows that initial expression of the *lac* operon genes promotes further expression in an autocatalytic manner due to positive feedback loop generated by the action of the permease. This positive genetic architecture is the cause of the experimentally observed all-or-none bistable response of the *lac* operon (Novick and Weiner 1957).

In order to reveal the role of the *lac* operon components, several experimental studies have introduced mutations to the promoter region or the coding sequences. Such mutations can result in strengthening, weakening or completely eliminating some of the biomolecular interactions in the *lac* operon system. Thus, the negative feedback of glucose by the catabolite repression mechanism can be nullified by using a mutant promoter for the *lacZYA*, such as the *lac* UV5 promoter (Reznikoff and Abelson 1978). On the other hand, the strength of repression can be altered by: (i) introducing mutations
to the lacI promoter, resulting in different transcriptional rate of lacI (Calos 1978; Calos and Miller 1980) or (ii) changing the binding affinities of LacI to the lacO or to the inducer IPTG by introducing mutations to the lacI gene (Markiewicz et al. 1994; Suckow et al. 1996; Falcon and Matthews 1999; Swint-Kruse et al. 2003). Moreover, the autocatalytic feedback can be suppressed by introducing mutations to the lacY gene, such that non-functional permease is produced (Padan et al. 1985).

4.2 Modeling Studies

The processes that take place during the expression of lac operon genes and the effect of mutations can be better understood by the use of mathematical modeling in close association with the experiments. Therefore, soon after the introduction of the operon concept by Jacob et al. (1960), modeling studies of the lac operon appeared in the literature.

By using a generic model for single gene induction, Griffith (1968) showed that in a positive feedback loop architecture, such as that exhibited in the lac operon, bistability is possible if more than one inducer molecules interact with the genetic locus. In Yagil and Yagil (Yagil and Yagil 1971), a reaction scheme was proposed for the induction mechanism present in the lac operon, and it was shown experimentally that the Hill exponents for the induction of LacI have values close to 2, implying that two inducer molecules are required for this process.

Several subsequent models (Van Dedem and Moo-Young 1973; Toda 1976; Imanaka and Aiba 1977; Gondo et al. 1978; Vieth et al. 1982) utilized the reaction scheme by Yagil and Yagil (Yagil and Yagil 1971) to model induction and catabolite repression in the case of multiple substrates, in different growth environments or by taking into
account membrane transport. In two milestone papers, Lee and Bailey (1984b; 1984a) derived a detailed model which enabled them to investigate replication dynamics and polyploidy effects, and their coupling to the induction-repression mechanisms, transcription and translation. Laffend and Schuler extended the work of Lee and Bailey and showed that for high copy plasmids, the ribosomes are not sufficient to translate the plasmid-derived mRNA into proteins (Laffend 1991; Laffend and Shuler 1994b; 1994a).

Furthermore, Ray et al. (1987) presented the first unified lac operon model that incorporated the role of the catabolite modulator factor in the regulation of transcription. Chung and Stephanopoulos (1996) presented a minimal, yet elegant, lac operon model, which takes into account repression and induction, as well as the positive feedback generated by the LacY permease. Straight and Ramkrishna (1991) and Ramakrishna et al. (1996), applied the cybernetic modeling approach and successfully predicted diauxic growth patterns and simultaneous consumption of substrates. Moreover, Wong et al. (1997) developed a detailed mathematical model to study inducer exclusion, catabolite repression and diauxic growth. Recently, Tian and Burrage (2005) constructed a model that accounts for transcription and translation of the LacY permease and a reporter gene GFP, and quantified the effect of different Hill coefficients for the transport of TMG on the LacY expression levels.

A class of models that take into account transcriptional and translational delays has also been developed. Maffahy and Savev (1999) showed that oscillatory behavior is possible in the lac operon for specific values of the transcriptional and translational delays. Yildirim and Mackey (2003) developed a detailed lac operon model consisting of delay differential equations and successfully predicted the experimentally observed time
course of β-galactosidase concentration and the bistable response of the lac operon.

In addition, Tanaka et al. (2006) have developed a generic framework for transcripational regulation that can also be used in the case of the lac operon system. The authors show that several published lac operon models can be reproduced by choosing the appropriate functional expressions in their framework.

Finally, Dreisigmeyer et al. (2008) developed a lac operon model in order to identify biologically meaningful parameter regimes where the system can exhibit bistability. The authors report that, in the case where artificial inducers are used, the system exhibits bistability over a range that can be controlled by model parameters. On the other hand, induction by the natural inducer (lactose) results in no bistability. Based on this observation, the authors claim that bistability is irrelevant in the naturally occurring lac operon. In our study we will model an engineered lac operon system induced by IPTG, an artificial analog of lactose, and thus we will observe bistability.

None of the aforementioned works takes into account the inherent randomness of the biomolecular mechanisms of the lac operon system. However, there have been published studies which incorporate stochastic effects on the mathematical description of the lac operon. Carrier and Keasling (1997) used a unique mechanistic model to simulate the processes involved in the lacZ gene expression focusing on the effects of particular mRNA decay mechanisms. An adaptation of this model was later used by the same authors (Carrier and Keasling 1999) in order to study the properties of gene expression in autocatalytic systems. It was shown that random cellular events govern the all-or-none phenomenon upon induction and that hysteretic responses can be observed. The authors also demonstrated that for homogeneous expression of the gene of interest, the inducer
must not control the transport protein levels.

Vilar et al. (2003) derived a phenomenological deterministic model for the induction of the *lac* operon which predicts bistability and hysteresis. They incorporated stochasticity by converting the deterministic reaction terms into probability transition rates and the concentrations into numbers of molecules. The stochastic model predicts random switches from the uninduced to the induced state. However, as the authors note, the model predicts that all cells eventually reach the induced state, unless one considers different growth rates for the induced versus the uninduced cells. Vilar et al. (2003) pointed out that taking into account stochastic and population effects depends not only on the given system but also on the particular conditions.

Mettetal et al. (2006) started from a deterministic *lac* operon dynamic model (developed by Ozbudak et al. 2004) into which white noise was added *a posteriori* ("Langevin approach", van Kampen 1992). The parameter values for the deterministic model were taken such that the model predictions agree with the dynamics of the experimentally measured cell population average. Furthermore, measurements of the noise around the steady state were used to estimate the parameters that determine the variance of the white noise terms. This model deals collectively with global noise, and also takes into account fluctuations in the GFP and Red Fluorescent Protein (RFP) concentrations. Mettetal et al. (2006) showed that even though the deterministic model cannot fully capture the experimental observations, the stochastic model correctly does so, after the noise characteristics have been adjusted to agree with the data.

Finally, a recent stochastic model by Saiz and Vilar (Saiz and Vilar 2006) focuses on the assembly of the LacI repressor complex with the operator sequence. The authors use
this model to predict transcription rates of the Lac-repressed gene and to also investigate the effect of mutations to the repressor, or the presence of pseudo-operators.

4.3 Conclusion

In this chapter we described the underlying biomolecular mechanisms present in the lac operon and reviewed the pertinent experimental literature. We further reviewed the deterministic and stochastic modeling studies for this system. The review of modeling studies reveals that while several deterministic models have successfully captured the salient features of the system’s behavior, little attention has been paid to the potential effects of stochasticity. Thus, in the following chapter we will formulate deterministic and stochastic single cell models for the lac operon network. By comparing the two models we will quantify the effect of intrinsic stochasticity (randomness in the occurrence of intracellular reaction events) for this system at the single cell level.
Chapter 5:
Quantifying the Effect of Intrinsic Noise for the \textit{lac} operon
Network at the Single Cell Level

In this chapter, we will assess the effect of stochasticity in intracellular reaction occurrence (intrinsic noise) on the behavior of the \textit{lac} operon system. To this end, we will first develop a reaction network that captures the salient features of the underlying interactions. Subsequently, we will derive stochastic and deterministic models from that network and, by comparing the two, we will assess the effect of intrinsic noise.

5.1 Motivation

As concluded in the previous chapter, no study has compared corresponding deterministic and stochastic models in order to show whether the extra complexity reveals novel phenomena in the case of the \textit{lac} operon system. Such phenomena have been demonstrated for other genetic networks, such as regulatory network motifs, where a protein produced by transcriptional-translational bursts exerts positive or negative feedback to other promoters (McAdams and Arkin 1997; McAdams and Arkin 1998). In these systems, the random occurrence of biomolecular reactions results in temporal
variability in the numbers of proteins. Consequently, phenotypic variability is observed even when the average protein numbers stay constant (McAdams and Arkin 1997). Further, it has been proposed that this variability may have a physiological significance, enhancing the viability of the cell population during environmental stress conditions or establishing cell population heterogeneity during cellular differentiation and development (Kærn et al. 2005; Veening et al. 2008a).

Phenotypic variability resulting from stochasticity has also been demonstrated in more complex systems. For the phage λ lysis-lysogeny decision circuit, Arkin et al. (1998) showed that the inherent randomness of the biomolecular events coupled with the system’s bistable response results in switching between different dynamical states. Consequently, the bacterial population splits into subpopulations of infected and uninfected cells. Deterministic models cannot account for this probabilistic switching, which is a key mechanism in the selection between alternative regulatory paths. For the same system, Hasty et al. (2000) also showed that adding external noise into a deterministic model can induce switching or amplification effects.

The profound effect of stochasticity in _de novo_ creating dynamical responses that cannot be predicted by deterministic modeling has been demonstrated in enzymatic futile cycles. Samoilov et al. (2005) showed that in the absence of stochasticity the system is unable to exhibit bistable behavior, but when noise is added, stochastic switching (bistability) is observed.

Furthermore, Kepler and Elston (2001) have developed exact stochastic models for several different genetic architectures, namely single gene without feedback, single gene activating itself, and two genes mutually repressing each other. They have further derived
approximations valid for low noise strength or rapid stochastic fluctuations and by using
the resulting approximate models, they were able to demonstrate bifurcations in the
stochastic models, even when the corresponding deterministic equations remain
unchanged.

In the current study, we apply a different approach for the case of the *lac* operon
system. Starting from the deterministic model, we compare its temporal asymptotic
behavior with that of the corresponding exact stochastic model in order to show the effect
of stochasticity. To this end, we first develop a reaction network that incorporates
biological information about the *lac* operon. We derive a deterministic and the
Corresponding stochastic model and analyze quantitatively the effect of biomolecular
parameters on the behavior of each model. We further compare the predictions of the two
models revealing mechanisms that significantly influence the temporal asymptotic
behavior of the stochastic, but not that of the deterministic model. Since no
approximations are imposed on either model, other than those used in building the
reaction network, this comparative approach can highlight the differences between the
two models thereby elucidating the effect of stochasticity on the behavior of the *lac*
operon system.

In the rest of the chapter, we first discuss the assumptions used to build the *lac* operon
reaction network. Subsequently, we derive the deterministic model, analyze its temporal
asymptotic behavior and perform sensitivity analysis with respect to parameters
quantifying key biomolecular mechanisms. We further derive the stochastic model and
use Gillespie's algorithm to simulate the stochastic dynamics (Gillespie 1977). We
present an in-depth analysis of the behavior of the system at different induction levels
highlighting the differences between the corresponding models. Finally, we summarize and discuss our results.

5.2 Reaction Network

The interacting species taken into account in the *lac* operon chemical reaction network appear in Table 5.1 and the molecular mechanisms included in our model, appear schematically in Figure 5.1. The constitutive expression of the *lacI* gene is modeled as 0\textsuperscript{th} order production of *lacI* mRNA (species M\textsubscript{R}). The translation of *lacI* mRNA to LacI (species R) is modeled as a 1\textsuperscript{st} order catalytic reaction and the dimerization of LacI repressor (to form the dimer R\textsubscript{2}) is assumed to follow 2\textsuperscript{nd} order kinetics. Thus, production and dimerization of LacI is modeled by the reactions:

\[
\begin{align*}
\varnothing & \xrightarrow{k_{\text{bir}}} M_R \\
M_R & \xrightarrow{k_{a}} M_R + R \\
2R & \xleftarrow{k_{-2r}} R_2
\end{align*}
\]  

(5.1, 5.2, 5.3, 4)

Wild type LacI dimers can further assemble into a tethered tetramer (Schmitz et al.

<table>
<thead>
<tr>
<th>Table 5.1: <em>lac</em> operon species notation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Symbol</strong></td>
</tr>
<tr>
<td>(M_R)</td>
</tr>
<tr>
<td>(R)</td>
</tr>
<tr>
<td>(R_2)</td>
</tr>
<tr>
<td>(O)</td>
</tr>
<tr>
<td>(R_2O)</td>
</tr>
<tr>
<td>(I)</td>
</tr>
<tr>
<td>(I_{ex})</td>
</tr>
<tr>
<td>(I_2R_2)</td>
</tr>
<tr>
<td>(M_Y)</td>
</tr>
<tr>
<td>(Y)</td>
</tr>
<tr>
<td>(Y_{ex})</td>
</tr>
<tr>
<td>(\varnothing)</td>
</tr>
</tbody>
</table>
The tetramer can interact with pseudo-operators and create loop structures (Matthews 1992) which result in tighter repression due to cooperativity between the (pseudo)operators (Oehler et al. 1994). However, in this study we are not concerned with the effects of cooperativity in the negative feedback due to repression. We rather aim in understanding the effect of stochasticity in presence of positive feedback due to autocatalytic LacY production. Thus, we neglect tetramerization effects, assuming mutant repressor which can form dimers but is unable to assemble into the tetrameric structure. Such mutants have been experimentally constructed (Schmitz et al. 1976; Chen and Matthews 1992; Chen et al. 1994b) and are unable to create loop structures (Matthews 1992).

A LacI dimer is sufficient for specific binding to the operator sequence (Lewis et al. 1996). However, a LacI monomer is unable to bind lacO and exert any repressive effect (Schmitz et al. 1976). Thus, we assume that the repressor dimer binds to the operator (denoted as species O) with a one-to-one stoichiometry. Under these assumptions, repressive action is expressed as:

\[
R_2 + O \stackrel{k_r}{\rightleftharpoons} R_2O
\]  

(5.5, 6)
Furthermore, induction of the lac operon is assumed to take place with a gratuitous inducer, such as TMG or IPTG (denoted as species I), which does not require transformation by the β-galactosidase. Consequently, the dynamics of this enzyme are not taken into account. Equilibrium dialysis studies have shown that IPTG binds strongly to the LacI tetramer with a stoichiometry that varies depending on the conditions, but reaches a value of 4 at high temperatures (one IPTG molecule per LacI monomer, Ohshima et al. 1974). In a LacI tetramer, the inducer binding to each LacI dimer can be considered separately, because the only allosteric changes in the repressor molecule occur between chains in a dimer (Bell and Lewis 2000). Hence, the interaction of a LacI dimer with two IPTG molecules is adequate in describing the mechanism of allosteric regulation, and thus we use this mechanism in our model. We assume cooperative binding, taking the Hill coefficient equal to 2 for the binding of inducer to repressor. Therefore, the phenomenon can be modeled as a reversible 3rd order reaction:

\[ 2I + R_2 \xrightleftharpoons[k_{-a1}]{k_{a1}} I_2R_2 \] (5.7, 8)

where (5.7) is the forward reaction. Furthermore, IPTG can also bind to the repressor-operator complex, thereby freeing the operator from the repressor:

\[ 2I + R_2O \xrightleftharpoons[k_{a2}]{k_{-a2}} I_2R_2 + O \] (5.9, 10)

Experimental studies have shown different Hill coefficient values, ranging from one to three (Yagil and Yagil 1971). Cooperativity values depend on whether the inducer binds free repressor versus repressor-operator complex, can be altered by mutations and depend on environmental conditions such as pH (Yagil and Yagil 1971; O’Gorman et al. 1980; Chen et al. 1994a). However, in our study we are not concerned with such structures, and
thus we make the simplifying assumption that cooperativity is equal to two as suggested by the reaction scheme of Yagil and Yagil (Yagil and Yagil 1971).

Note that the reversible reactions (5.5, 6), (5.7, 8) and (5.9, 10) are linearly dependent; consequently, the following equation for their dissociation constants holds:

\[
\frac{k_{dr1} k_{dr2}}{k_{r} k_{dr1} k_{dr2}} = 1
\]  

(5.11)

Equation (5.11) essentially expresses Hess’s law for the Gibbs free energy applied to the set of chemical equilibria (5.5 - 9) (see Appendix III, page 168).

For lacY transcription, we use a catalytic 1st order reaction where the rate of lacY mRNA (species \( M_Y \)) production depends on the concentration of the free operator. We assume that the transcription of lacY is driven by a mutant promoter insensitive to glucose, such as the UV5, and thus the catabolite repression effects are neglected:

\[
O \xrightarrow{k_{\text{unr}}} O + M_Y
\]  

(5.12)

Leak transcription of lacY is modeled by the following reaction:

\[
R_2O \xrightarrow{k_{\text{unr}}} R_2O + M_Y
\]  

(5.13)

Translation of lacY gene to LacY protein (species \( Y \)) is modeled as a 1st order catalytic reaction:

\[
M_Y \xrightarrow{k_{\text{tr}}} M_Y + Y
\]  

(5.14)

The facilitated transport of IPTG from the extracellular to the cytosolic space was assumed to follow Michaelis-Menten kinetics (Kepes 1960). The LacY permease plays the role of the enzyme and the extracellular IPTG (\( I_{\text{ex}} \)) that of the substrate:

\[
Y + I_{\text{ex}} \xrightarrow{k_p} YI_{\text{ex}} \xrightarrow{k_{\text{r}}} Y + I
\]  

(5.15 - 17)

where (5.15) is the reaction \( k_p \) and (5.17) corresponds to reaction \( k_{\text{r}} \). Moreover, the free
diffusion of IPTG through the membrane was described by two first order reactions with the exact same kinetic constants:

\[
I_{\text{ex}} \xrightleftharpoons[\dot{k}_t]{k_i} I \tag{5.18, 19}
\]

where \( k_i = k_{-1} = \frac{h_i \cdot A_{E.coli}}{V_{E.coli}} \) (see Appendix VI, page 178). The value of \( k_i \) can be calculated if the transport coefficient \( h_i \), membrane area and cell volume are known.

Degradation reactions were modeled using the following assumptions: since the operator does not degrade, neither does the operator-repressor complex (the repressor has to dissociate first and then degrade). Also, IPTG is non-metabolizable, so it does not degrade itself, but the repressor-inducer complex degrades to give two IPTG molecules. Similarly, the permease inducer complex degrades in the interior of the cell to give one IPTG molecule. We further assume that the repressor does not degrade when bound to the operator. We have not found experimental data supporting this assumption, but one expects that the chaperones/proteases would not be able to access the repressor when bound to DNA due to steric hindrance effects. Last, all degradation reactions are assumed to follow first order kinetics:

\[
M_R \xrightarrow{\lambda_{\text{de }}} \emptyset \tag{5.20}
\]

\[
M_Y \xrightarrow{\lambda_{\text{de }}} \emptyset \tag{5.21}
\]

\[
R \xrightarrow{\lambda_R} \emptyset \tag{5.22}
\]

\[
R_2 \xrightarrow{\lambda_{R2}} \emptyset \tag{5.23}
\]

\[
Y \xrightarrow{\lambda_Y} \emptyset \tag{5.24}
\]

\[
YI_{\text{ex}} \xrightarrow{\lambda_{\text{de}}} I \tag{5.25}
\]
\begin{equation}
I_2R_2 \xrightarrow{\lambda_{HB2}} 2I
\end{equation}

We further neglected cell growth and division, assuming that the total DNA and therefore the operator concentration in the cell remain constant, and that the cell volume remains constant. This simplifying assumption allows us to analyze in isolation the genuine effects of stochasticity in biochemical reaction occurrence. Cell growth and division make significant contributions to the total reaction network constituting the cell. Thus, the cell growth and division process should be

Table 5.2: Parameters of the \textit{lac} operon reaction model

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Value</th>
<th>Units</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$V_{E.\ coli}$</td>
<td>(8 \cdot 10^{-16})</td>
<td>L</td>
<td>\textit{E. coli} volume (^1)</td>
</tr>
<tr>
<td>A</td>
<td>(4.75 \cdot 10^{-10})</td>
<td>dm(^2)</td>
<td>\textit{E. coli} membrane surface area (see eq. 2.25)</td>
</tr>
<tr>
<td>$O_T$</td>
<td>1</td>
<td>(copy number)</td>
<td>operator molecular content ([O] (\approx 2.08) nM)</td>
</tr>
<tr>
<td>$k_{sAMR}$</td>
<td>0.23</td>
<td>nM-min(^{-1})</td>
<td>\textit{lacI} transcription rate (^7)</td>
</tr>
<tr>
<td>$k_{sR}$</td>
<td>15</td>
<td>min(^{-1})</td>
<td>\textit{LacI} monomer translation rate constant (^7)</td>
</tr>
<tr>
<td>$k_{2R}$</td>
<td>50</td>
<td>nM(^{-1})-min(^{-1})</td>
<td>\textit{LacI} dimerization rate constant (^3)</td>
</tr>
<tr>
<td>$k_{-2R}$</td>
<td>(10^{-3})</td>
<td>min(^{-1})</td>
<td>\textit{LacI} dimer dissociation rate constant (^3)</td>
</tr>
<tr>
<td>$k_r$</td>
<td>960</td>
<td>nM(^{-1})-min(^{-1})</td>
<td>association rate constant for repression (^2)</td>
</tr>
<tr>
<td>$k_{-r}$</td>
<td>2.4</td>
<td>min(^{-1})</td>
<td>dissociation rate constant for repression (^2)</td>
</tr>
<tr>
<td>$k_{d1}$</td>
<td>(3 \cdot 10^{-7})</td>
<td>nM(^{-2})-min(^{-1})</td>
<td>association rate constant for 1(^{st}) derepression mechanism (^4)</td>
</tr>
<tr>
<td>$k_{-d1}$</td>
<td>12</td>
<td>min(^{-1})</td>
<td>dissociation rate constant for 1(^{st}) derepression mechanism (^4)</td>
</tr>
<tr>
<td>$k_{d2}$</td>
<td>(3 \cdot 10^{-7})</td>
<td>nM(^{-2})-min(^{-1})</td>
<td>association rate constant for 2(^{nd}) derepression mechanism (^5)</td>
</tr>
<tr>
<td>$k_{-d2}$</td>
<td>(4.8 \cdot 10^{3})</td>
<td>nM(^{-1})-min(^{-1})</td>
<td>dissociation rate constant for 2(^{nd}) derepression mechanism (^5)</td>
</tr>
<tr>
<td>$k_{sAMY}$</td>
<td>0.5</td>
<td>min(^{-1})</td>
<td>\textit{lacY} transcription rate constant (^7)</td>
</tr>
<tr>
<td>$k_{sMY}$</td>
<td>0.01</td>
<td>min(^{-1})</td>
<td>leak \textit{lacY} transcription rate constant (^7)</td>
</tr>
<tr>
<td>$k_{sY}$</td>
<td>30</td>
<td>min(^{-1})</td>
<td>\textit{lacY} translation rate constant (^7)</td>
</tr>
<tr>
<td>$k_p$</td>
<td>0.12</td>
<td>nM(^{-1})-min(^{-1})</td>
<td>LacY-inducer association rate constant (^6)</td>
</tr>
<tr>
<td>$k_{-p}$</td>
<td>0.1</td>
<td>min(^{-1})</td>
<td>LacY-inducer dissociation rate constant (^6)</td>
</tr>
<tr>
<td>$\lambda_r$</td>
<td>(6 \cdot 10^4)</td>
<td>min(^{-1})</td>
<td>IPTG facilitated transport constant (^6)</td>
</tr>
<tr>
<td>$\beta_c$</td>
<td>(1.55 \cdot 10^{-6})</td>
<td>dm-min(^{-1})</td>
<td>IPTG passive diffusion permeability constant (^6)</td>
</tr>
<tr>
<td>$\lambda_{AMR}$</td>
<td>0.462</td>
<td>min(^{-1})</td>
<td>\textit{lacI} mRNA degradation constant (^7)</td>
</tr>
<tr>
<td>$\lambda_{MY}$</td>
<td>0.462</td>
<td>min(^{-1})</td>
<td>\textit{lacY} mRNA degradation constant (^7)</td>
</tr>
<tr>
<td>$\lambda_r$</td>
<td>0.2</td>
<td>min(^{-1})</td>
<td>\textit{LacI} monomer degradation constant (^7)</td>
</tr>
<tr>
<td>$\lambda_r$</td>
<td>0.2</td>
<td>min(^{-1})</td>
<td>\textit{LacI} dimer degradation constant (^7)</td>
</tr>
<tr>
<td>$\lambda_Y$</td>
<td>0.2</td>
<td>min(^{-1})</td>
<td>\textit{LacY} degradation constant (^7)</td>
</tr>
<tr>
<td>$\lambda_{Y\text{lex}}$</td>
<td>0.2</td>
<td>min(^{-1})</td>
<td>\textit{LacY}-inducer degradation constant (^7)</td>
</tr>
<tr>
<td>$\lambda_{D\gamma 2}$</td>
<td>0.2</td>
<td>min(^{-1})</td>
<td>repressor-inducer degradation constant (^7)</td>
</tr>
</tbody>
</table>
1. *E. coli* volume is roughly the average between $6 \cdot 10^{-16}$ and $9.8 \cdot 10^{-16}$ (Santillán and Mackey 2004).

2. The equilibrium constant $k_\text{r}/k_\text{i}$ is $10^{-13} \sim 10^{-11}$ M, (Riggs et al. 1970b; Riggs et al. 1970a; Goeddel et al. 1977). The half-life for dissociation of operator DNA fragments from repressor, has been reported close to 30 to 49 s (Lin et al. 1976; Goeddel et al. 1977). We take the dissociation rate $k_\text{r} = 2.4 \text{ min}^{-1}$, $k_\text{i} = 960 \text{ nM}^{-1}\text{min}^{-1}$ (thus $k_\text{r}/k_\text{i} = 2.5 \cdot 10^{-12}$).

3. Experimental data for the dimerization rate or the thermodynamic constant could not be found; thus, these values are estimated and express fast and tight dimerization.

4. The equilibrium constant $k_\text{d}^{-1}/k_\text{d}$ is $4 \cdot 10^{-12} \text{ M}^2 = 4 \cdot 10^7 \text{ nM}^2$ (Yagil and Yagil 1971). The dissociation rate $k_\text{d}^{-1}$ is 0.2 s$^{-1} = 12 \text{ min}^{-1}$ (Dunaway et al. 1980).

5. For the association constant for the second derepression mechanism ($k_\text{d}^{-1}$) we used a value equal to that of the first mechanism ($k_\text{d}$). Parameter $k_\text{d}^{-1}$ was calculated using equation (5.11).

6. For the Michaelis-Menten scheme, $(k_\text{r} + k_p)/k_p = 5 \cdot 10^5 \text{ nM}$, $k_p = 6 \cdot 10^4 \text{ min}^{-1}$ and $k_i$ reported for TMG is $0.92 \text{ min}^{-1}$ (Cohn and Horibata 1959; Kepes 1960; Chung and Stephanopoulos 1996). Thus, we use $k_p = 0.1 \text{ min}^{-1}$, $k_i = 0.12 \text{ nM}^{-1}\text{min}^{-1}$ (see also eq. 5.18, 19).

7. Parameters for which experimental values could not be determined from the literature were estimated so that the species’ concentrations were within reasonable limits. Thus, (a) the ratio of the intracellular with respect to the extracellular IPTG concentration must be to the order of 26 ~ 400 (Kepes 1960) and (b) at the fully induced state there are roughly 400 ± 200 nM LacY permease or equivalently 200 ± 100 LacY permease molecules (Cohn and Horibata 1959). We used this information to estimate the production and degradation rates of $M_Y$ and $Y$. Furthermore, (c) the total concentration of dimeric repressor in the cell is estimated to be 20 nM or equivalently 10 repressor dimers (Gilbert and Müller-Hill 1966; Yagil and Yagil 1971); thus, we estimated production and degradation constants for the species $M_R$, $R$, $I_R$ and $I_R^2$ that agree with this data. The order of magnitude for the production and degradation rates was taken such that upon step changes of $[I]_\text{on}$ the system will respond within a reasonable *E. coli* division time (~ 40 min maximum).

8. Parameters for experimental values could not be determined from the literature were estimated so that the species’ concentrations were within reasonable limits. Thus, (a) the ratio of the intracellular with respect to the extracellular IPTG concentration must be to the order of 26 ~ 400 (Kepes 1960) and (b) at the fully induced state there are roughly 400 ± 200 nM LacY permease or equivalently 200 ± 100 LacY permease molecules (Cohn and Horibata 1959). We used this information to estimate the production and degradation rates of $M_Y$ and $Y$. Furthermore, (c) the total concentration of dimeric repressor in the cell is estimated to be 20 nM or equivalently 10 repressor dimers (Gilbert and Müller-Hill 1966; Yagil and Yagil 1971); thus, we estimated production and degradation constants for the species $M_R$, $R$, $I_R$ and $I_R^2$ that agree with this data. The order of magnitude for the production and degradation rates was taken such that upon step changes of $[I]_\text{on}$ the system will respond within a reasonable *E. coli* division time (~ 40 min maximum).

In summary, the reaction network from which the two corresponding models will be derived consists of reactions (5.1 - 10) and (5.12 - 26). Species notation is summarized in Table 5.1 and the values of the kinetic constants and other parameters are shown in Table 5.2. To obtain a better understanding of the model and to gain insight into the effect of various mechanisms on its behavior, we will first analyze the deterministic equations arising from the reaction network and then investigate the effect of stochasticity.

### 5.3 Deterministic Model

The deterministic model consists of the mass balances for the interacting species, subject to conservation conditions. Concentration of species X is denoted as $[X]$ (for species notation, see Table 5.1). The total operator concentration, denoted $[O]_T$, is
constant and equal to the free operator concentration plus that bound to the repressor.

\[ [O]_T = [O] + [R_2O] \]  

(5.27)

Utilizing the above relation, the mass balances are written as:

\[ \frac{d[M_{R2}]}{dt} = k_{sMR} - \lambda_{sMR} \cdot [M_R] \]  

(5.28)

\[ \frac{d[R]}{dt} = k_{sR} \cdot [M_R] - 2 \cdot k_{2R} \cdot [R]^2 + 2 \cdot k_{-2R} \cdot [R_2] - \lambda_{R} \cdot [R] \]  

(5.29)

\[ \frac{d[R_2]}{dt} = k_{2R} \cdot [R]^2 - k_{-2R} \cdot [R_2] - k_r \cdot [R_2] \cdot [O] + k_{dr1} \cdot ([O]_T - [O]) \] 

\[-k_{dr1} \cdot [R_2] \cdot [I]^2 + k_{-dr1} \cdot [I_2R_2] - \lambda_{R2} \cdot [R_2] \]  

(5.30)

\[ \frac{d[O]}{dt} = -k_r \cdot [R_2] \cdot [O] + k_{-r} \cdot ([O]_T - [O]) + k_{dr2} \cdot ([O]_T - [O]) \cdot [I]^2 \] 

\[-k_{-dr2} \cdot [O] \cdot [I_2R_2] \]  

(5.31)

\[ \frac{d[I]}{dt} = 2 \cdot k_{d1} \cdot [R_1] \cdot [I]^2 + 2 \cdot k_{-d1} \cdot [I_2R_2] - 2 \cdot k_{dr2} \cdot ([O]_T - [O]) \cdot [I]^2 \] 

\[+2 \cdot k_{-dr2} \cdot [O] \cdot [I_2R_2] + k_h \cdot [Y_{Iex}] + h_i \cdot \frac{A}{V} \cdot ([I_{ex}] - [I]) \] 

\[+2 \cdot \lambda_{I2R2} \cdot [I_2R_2] + \lambda_{Y_{Iex}} \cdot [Y_{Iex}] \]  

(5.32)

\[ \frac{d[I_2R_2]}{dt} = k_{dr1} \cdot [R_2] \cdot [I]^2 - k_{-dr1} \cdot [I_2R_2] + k_{dr2} \cdot ([O]_T - [O]) \cdot [I]^2 \] 

\[-k_{-dr2} \cdot [O] \cdot [I_2R_2] - \lambda_{I2R2} \cdot [I_2R_2] \]  

(5.33)

\[ \frac{d[M_Y]}{dt} = k_{s0MY} \cdot ([O]_T - [O]) + k_{sMY} \cdot [O] - \lambda_{MY} \cdot [M_Y] \]  

(5.34)

\[ \frac{d[Y]}{dt} = k_{sY} \cdot [M_Y] + (k_{hi} + k_{-p}) \cdot [Y_{Iex}] - k_p \cdot [Y] \cdot [I_{ex}] - \lambda_Y \cdot [Y] \]  

(5.35)

\[ \frac{d[Y_{Iex}]}{dt} = -(k_{hi} + k_{-p}) \cdot [Y_{Iex}] + k_p \cdot [Y] \cdot [I_{ex}] - \lambda_{Y_{Iex}} \cdot [Y_{Iex}] \]  

(5.36)
5.3.1 Temporal Asymptotic Behavior

In order to study the behavior of the lac operon as predicted by the deterministic model, we will first solve for the steady state concentrations of the species. The extracellular IPTG concentration \([I_{ex}]\) will be used as the main bifurcation parameter, since it can be experimentally varied, hence rendering the connection between theory and experiments possible. Furthermore, it will be discussed how key parameters linked to biomolecular mechanisms affect the system's behavior. The primary observable of the system will be the total LacY concentration (namely \([Y] + [YI_{ex}]\)), since it can be measured experimentally with the use of reporter proteins or Western blot analysis.

In Figure 5.2a, the steady state concentration of total LacY is plotted with respect to the extracellular inducer concentration \([I_{ex}]\). All one-parameter bifurcation diagrams were computed by performing pseudo-arc length continuation for the steady states of equations (5.28 - 36). We observe that for low or high IPTG concentrations, only one stable steady state exists. Within a range of \([I_{ex}] \approx 24.2 - 32.4 \, \mu M\), two branches of stable steady states exist, separated by a branch of unstable steady states.

This bistable response can be explained in terms of the positive feedback architecture of the lac operon system. For low concentrations of the extracellular IPTG, \([I_{ex}]\), a minimal amount of LacY is produced, so the \(lacY\) gene essentially remains turned off (left monostable regime). For intermediate \([I_{ex}]\) (bistable region) if the initial content of LacY in the cell is low, then it will be impossible to turn on the \(lacY\) gene. However, if the initial LacY content is high enough, the \(lacY\) gene will be turned on and remain so, as a consequence of the facilitated inducer transport into the cytosol. Finally, for very high \([I_{ex}]\) the \(lacY\) gene cannot be turned off. Even if the initial LacY was zero, the IPTG that
freely diffuses in this case is sufficient to turn on the gene and activate the positive feedback loop. Thereafter, the facilitated transport due to the LacY permease dominates, the cell always reaches the induced state and bistability vanishes. Such threshold phenomena in the dynamics of induction, have also been observed experimentally (Novick and Weiner 1957; Cohn and Horibata 1959) and are commonly referred to as "all-or-none" phenomena.

The facilitated IPTG transport, occurring after induction, also generates a bistable response for the intracellular IPTG, as shown in Figure 5.2b. Moreover, the plot of
intracellular with respect to extracellular IPTG reveals a saturation effect: for high induction levels, the total LacY that exists in the cell remains mostly in the bounded form, YI_{ex}, (for a mathematical proof of this statement see Appendix IV, page 169). Thus, the remaining free LacY is not sufficient to transport IPTG into the cell and consequently no facilitated diffusion can take place. This explains the drop in the rate with which the intracellular IPTG concentration increases (after [I_{ex}] = 100 μM). On the other hand, for small induction levels, the intracellular and the extracellular IPTG will also be approximately equal since only a leak amount of LacY exists to facilitate the IPTG transport. Between the two limiting cases, there is a region where the intracellular IPTG concentration is roughly 40-fold higher than the extracellular. This value is within the range measured for other galactosides: 26 for TPG, 65 for TMG and 400 for TDG (Kepes 1960), and it can be increased by (results not shown): (a) faster LacY production or higher total operator concentrations [O]_T, (b) faster association of the extracellular to the LacY permease (higher values for k_p) and (c) faster facilitated transport (higher values for k_R).

Furthermore, the total repressor dimer concentration [R_2]_T = [R_2] + [R_2O] + [I_2R_2] also assumes limiting values for low and high induction levels (Figure 5.2c). Between the two limits (low and high induction) there exists a region where bistability is observed for the total [LacI]. The prediction of this bistable response for [R]_T is a novel feature of our model, which results from the incorporation of the LacI dynamics into the reaction network. Previous models neglect the LacI dynamics and assume a non-regenerative repressor pool ([R_2]_T = const.), which is used by the cell to block lacY transcription. When IPTG is added, it binds with a high affinity to the repressor, brings about the
depletion of the available free LacI and results in the cessation of lacY suppression. Therefore, the “switch-on” mechanism in those models is based on the depletion of LacI from the non-regenerative repressor pool. On the contrary, in our model we have relaxed the assumption of constant total repressor concentration. In this case, higher induction levels result in higher repressor-inducer complex concentrations (I₂R₂) and lower repressor-operator complex concentrations (R₂O). Since both the free repressor and the complex degrade with the same rate, but the R₂O complex does not degrade at all, it follows that higher IPTG concentrations lead to a higher net loss of repressor. Thus, in our model the “switch on” mechanism is based on the differential degradation of the repressor-inducer and repressor-operator complexes.

Accordingly, Figure 5.3a reveals that the degradation rate of the repressor-inducer complex (λ₁R₂) is a key parameter affecting the region of the parameter space where bistability is observed. The plot presents a 2-parameter bifurcation diagram for the turning points with respect to the degradation rate (λ₁R₂) and the extracellular IPTG concentration. This and all subsequent two parameter bifurcation diagrams were computed using 0th order continuation to the formulation that characterizes the turning point bifurcation (equations 2.27 - 2.29 in Salinger et al. 2002). We observe that, as the repressor-inducer complex degradation rate (λ₁R₂) tends to zero, the bistable regime shifts progressively to larger extracellular IPTG concentrations. For the limiting case where the complex does not degrade at all, the system becomes non-inducible.

There are several other parameters that have a significant effect on the region of bistability and also can be experimentally manipulated in order to make possible the connection with experiments. Thus, in Figure 5.3b the effect of the lacI transcription rate
Figure 5.3: Two parameter bifurcation diagrams for the turning points showing the effect of several biomolecular parameters on the bistable regime (enclosed between the curves). Shown are the effects of repressor-inducer complex degradation rate (a), of lacI transcription rate (b), of repressor-inducer dissociation rate (c) and of total operator concentration (d). Parameters as in Table 5.2 except as noted.

(k_{sMR}) is shown. The bistable regime is enclosed in the spindle-shaped area and it is apparent that higher lacI transcription rates result in shifting the bistable regime to higher extracellular inducer concentrations until the system becomes so repressed that bistability is destroyed. Qualitatively identical is the effect of the translation rate k_{sR} (results not shown). This effect is due to the higher LacI concentrations brought about by higher transcription (or translation) rates. In this case, induction will need more IPTG to inactivate the repressor, and thus the switching from the uninduced to the induced state occurs at higher extracellular [IPTG].
The region of bistability is also greatly affected by the binding of the repressor to the inducer, as shown in Figure 5.3c. For higher values of the inducer unbinding constant \( (k_{\text{dr1}}) \), the bistable regime shifts towards higher extracellular IPTG concentrations. This effect results from the binding of inducer to the repressor being weaker, and thus the inductive action less potent. Consequently, the system needs higher inducer concentrations to reach the induced state. On the contrary, the binding of the repressor to the operator follows opposite trend: weaker repressor-operator binding (higher \( k_\text{r-t} \) values) result in less potent repressive action. Thus, the system can be induced with lower inducer concentrations and the bistable regime shifts to lower extracellular [IPTG] (results not shown). Both of these mechanisms can be experimentally manipulated by introducing mutations to the \( lacI \) gene that affect the binding affinities of LacI to \( lacO \) or IPTG.

It is interesting to note that the total LacY concentration at maximal induction is not affected by either of the aforementioned parameters (namely \( lacI \) transcription rate \( k_{\text{4MR}} \), LacI-IPTG complex degradation rate \( \lambda_{12R2} \) and binding affinities of repressor to inducer \( k_{\text{dr1}}/k_{\text{r1-d1}} \) or to operator \( k_\text{r}/k_\text{r-t} \)). In fact, the maximal total LacY concentration can be calculated as (see equation AIV.41):

\[
\lim_{[\text{Is}] \rightarrow \infty} [Y]_T = \frac{k_{\text{SY}} \cdot k_{\text{MY}}}{\lambda_{\text{Ylex}} \cdot \lambda_{\text{MY}}} [O]_T \tag{5.37}
\]

This shows the intuitively expected fact that the repressor just serves to turn on and off a gene, but does not affect transcription or translation rates of the fully induced state. On the other hand, the maximal \([Y]_T\) scales linearly with the total operator concentration \([O]_T\), a parameter directly related to the copy number of the genetic material. Furthermore, the total operator concentration strongly affects the extent of the bistable
region. As shown in Figure 5.3d, higher $[O]_T$ values result in initially wider and subsequently narrower bistable regions, shifted to smaller extracellular IPTG concentrations $[I_{ex}]$. These effects are observed because of the strengthening of the positive feedback loop.

5.3.2 Alternative Derepression Mechanisms

The results presented so far pertain to the case where both derepression mechanisms are included in the model (reversible reactions 5.7, 8 and 5.9, 10). It is of experimental interest to demonstrate possible differences in the behavior of the system arising if only one or the other mechanism is functional. Such effects were studied in the temporal asymptotic limit, as well as in the transient behavior, by keeping the thermodynamic constants fixed and changing the rates of association and dissociation. For clarity we will refer to the reactions (5.7, 8) as the mechanism in which the inducer binds to the free repressor, and to the reactions (5.9, 10) as the mechanism in which the inducer binds to the repressor-operator complex.

Thus, the mechanism in which the inducer binds to the repressor-operator complex was found to destroy bistability for slower repressor-operator association and dissociation rates as shown in Figure 5.4a which combines three bifurcation diagrams for $[Y]_T$ versus $[I_{ex}]$ for three different rates. Furthermore, for this case faster induction is observed transiently (Figure 5.4b) because once the inducer has freed the operator from the repressor (fast process) the free repressor can bind again to the operator with a lower rate. The lower this rate, the faster the induction. On the other hand, the mechanism in which the inducer binds to the free repressor does not result in changes of the bifurcation
Figure 5.4: Effect of the alternative derepression mechanisms transiently and asymptotically. Panels (a) and (b) correspond to the mechanism where inducer binds only to repressor-operator complex (thus, for the other mechanism, $k_{dr1}$ and $k_{dr1}$ are both zero): loss of bistability and faster induction in the case of slower repressor-operator association and dissociation rates. $[k_r \ k_{-r}] = 50$ denotes that both $k_r$ and $k_{-r}$ have values 50-fold lower than those of Table 5.2. Panels (c) and (d) correspond to the mechanism where inducer binds only to free repressor (thus, $k_{dr2}$ and $k_{dr2}$ are both zero): Shift of the bistable regime and slower induction for the case of slower repressor-inducer association and dissociation rates. The time course for $[k_{dr1} \ k_{dr1}] = 500$ finally approaches the lower steady state. Panels (e) and (f) correspond to the mechanism where inducer binds to repressor-operator complex (thus, $k_{dr1}$ and $k_{dr1}$ are both zero): Broadening of the bistable regime and slower induction for the case of slower repressor operator complex-inducer association and dissociation rates. Parameters as in Table 5.2 except as noted. For all induction transients the IC is the corresponding steady state for $[I_{ex}] = 0 \mu M$ and at $t = 0$ min an extracellular IPTG concentration $[I_{ex}] = 100 \mu M$ is imposed.
structure, but it results in slower induction since, in this case, the repressor must first
dissociate from the operator and then bind to the inducer (results not shown).

Furthermore, for slower repressor-inducer association and dissociation rates, the
mechanism in which the inducer binds to the free repressor results in a shifting of the
bistable regime to higher $I_{eq}$ concentrations and delayed induction (Figure 5.4c, d). Both
effects can be attributed to repressor-inducer complex degradation occurring faster than
repressor-inducer association. Thus, the system needs higher IPTG concentrations in
order to effectively suppress the repressive action.

Finally, for slower repressor operator complex-inducer association and dissociation
rates, the mechanism in which inducer binds to the repressor-operator complex results in
a broadening, but not significant shifting of the bistable regime, and also in slower
induction (Figure 5.4e, f). These effects may be attributed to the progressively more
significant contribution of $I_2R_2$ degradation. This $I_2R_2$ "sink" results in shifting the
equilibrium of reactions (5.9, 10) to the right, thereby enhancing the positive feedback
effect and broadening the bistable regime.

For the case that both mechanisms coexist, lower association and dissociation rates for
any equilibrium produces no appreciable effects, because the system can preferentially
use one of the two derepression mechanisms thereby making up for the lower rates of the
other.

5.4 Stochastic Model

The deterministic model gave significant insight into the behavior of the *lac* operon
system. However, there are indications that for this system stochasticity may be
considerable. In particular, some of the interacting species have low copy numbers: the
number of LacY permease molecules at the fully induced state are roughly 200 ± 100 (Cohn and Horibata 1959), the total repressor molecules in the cell are estimated to be 10 dimeric copies per gene (Gilbert and Müller-Hill 1966; Yagil and Yagil 1971), which can be increased 10 or 20 fold as a result of the $f$ mutation (Gilbert and Müller-Hill 1966; Müller-Hill et al. 1968), and the operator copy number is 1 (or 2 during DNA duplication) for a chromosomal lacO. In view of these experimental data, it is reasonable to study the effects of stochasticity on network behavior, by comparing the predictions of the deterministic model to those of the corresponding stochastic model.

The stochastic model derived from our reaction network, consists of the chemical M-Equation (Gillespie 1977):

$$
\frac{\partial P(x, t|x_0, t)}{\partial t} = \sum_{j=1}^{m} \left[ \alpha_j(x - v_j) \cdot P(x - v_j, t|x_0, t) - \alpha_j(x) \cdot P(x, t|x_0, t) \right]
$$  

(5.38)

For our system, the M-Equation is impossible to solve analytically. Thus, we used the Direct Method of Gillespie's algorithm (Gillespie 1976; Gillespie 1977) in order to simulate the stochastic dynamical behavior of the lac operon system. The state vector containing numbers of molecules for each species is $x = \{ M_R, R, R_2, O, R_2O, I, I_2R_2, M_Y, Y, YI_{ex} \}$, and we have $n = 10$ species participating in $m = 25$ reactions. The reactions' propensity functions, $\alpha_j(x) = c_j \cdot h_j(x)$, $j = 1, 2, \ldots, m$, can be calculated given the macroscopic kinetic constants of Table 5.2, and are presented in Table 5.3. The vectors $v_j$ denote the change in the number of molecules for each species, e.g. for reaction (5.9), $v_9 = \{0, 0, 0, 1, -1, -2, 1, 0, 0, 0\}$.

From the simulated sample paths, one can estimate the mean and standard deviation of the total LacY concentration $[Y]_T$ for a range of extracellular [IPTG] and compare them
Table 5.3: Propensity functions for the stochastic *lac* operon model

<table>
<thead>
<tr>
<th>Reaction</th>
<th>Propensity Function $^{1,2,3}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(5.1) $\emptyset \xrightarrow{k_{M_{R}}} M_{R}$</td>
<td>$\alpha_1 = \frac{V_{Ecoli}}{N_A} \cdot k_{M_{R}}$</td>
</tr>
<tr>
<td>(5.2) $M_{R} \xrightarrow{k_{SR}} M_{R} + R$</td>
<td>$\alpha_2 = k_{SR} \cdot M_{R}$</td>
</tr>
<tr>
<td>(5.3) $2R \xrightarrow{k_{SR}} R_2$</td>
<td>$\alpha_3 = \frac{k_{SR}}{V_{Ecoli}} \cdot R \cdot (R-1)$</td>
</tr>
<tr>
<td>(5.4) $R_2 \xrightarrow{k_{SR}} 2R$</td>
<td>$\alpha_4 = k_{2R} \cdot R_2$</td>
</tr>
<tr>
<td>(5.5) $R_2 + O \xrightarrow{k_{SR}} R_2O$</td>
<td>$\alpha_5 = \frac{k_{SR}}{V_{Ecoli}} \cdot R_2 \cdot O$</td>
</tr>
<tr>
<td>(5.6) $R_2O \xrightarrow{k_{SR}} R_2 + O$</td>
<td>$\alpha_6 = k_{sr} \cdot R_2O$</td>
</tr>
<tr>
<td>(5.7) $2I + R_2 \xrightarrow{k_{SR}} I_2R_2$</td>
<td>$\alpha_7 = \frac{k_{dI}}{(V_{Ecoli} \cdot N_A)^2} \cdot R_2 \cdot I \cdot (I-1)$</td>
</tr>
<tr>
<td>(5.8) $I_2R_2 \xrightarrow{k_{dI}} 2I + R_2$</td>
<td>$\alpha_8 = k_{dI} \cdot I_2R_2$</td>
</tr>
<tr>
<td>(5.9) $2I + R_2O \xrightarrow{k_{dI}} I_2R_2 + O$</td>
<td>$\alpha_9 = \frac{k_{dI}}{(V_{Ecoli} \cdot N_A)^2} \cdot R_2O \cdot I \cdot (I-1)$</td>
</tr>
<tr>
<td>(5.10) $I_2R_2 + O \xrightarrow{k_{dI}} 2I + R_2O$</td>
<td>$\alpha_{10} = \frac{k_{dI}}{V_{Ecoli} \cdot N_A} \cdot I_2R_2 \cdot O$</td>
</tr>
<tr>
<td>(5.12) $O \xrightarrow{k_{M_{Y}}} O + M_{Y}$</td>
<td>$\alpha_{11} = k_{M_{Y}} \cdot O$</td>
</tr>
<tr>
<td>(5.13) $R_2O \xrightarrow{k_{M_{Y}}} R_2O + M_{Y}$</td>
<td>$\alpha_{12} = k_{M_{Y}} \cdot R_2O$</td>
</tr>
<tr>
<td>(5.14) $M_{Y} \xrightarrow{k_{B}} M_{Y} + Y$</td>
<td>$\alpha_{13} = k_{Y} \cdot M_{Y}$</td>
</tr>
<tr>
<td>(5.15) $Y + I_{ex} \xrightarrow{k_{i}} YI_{ex}$</td>
<td>$\alpha_{14} = k_{i} \cdot [I_{ex}] \cdot Y$</td>
</tr>
<tr>
<td>(5.17) $YI_{ex} \xrightarrow{k_{i}} Y + I_{ex}$</td>
<td>$\alpha_{15} = k_{i} \cdot YI_{ex}$</td>
</tr>
<tr>
<td>(5.18) $I_{ex} \xrightarrow{k_{i}} I$</td>
<td>$\alpha_{16} = k_{i} \cdot YI_{ex}$</td>
</tr>
<tr>
<td>(5.19) $I \xrightarrow{k_{i}} I_{ex}$</td>
<td>$\alpha_{17} = \frac{A_{Ecoli}}{V_{Ecoli}} \cdot h_i \cdot [I_{ex}]$</td>
</tr>
<tr>
<td>(5.20) $M_{R} \xrightarrow{k_{M_{R}}} \emptyset$</td>
<td>$\alpha_{18} = \frac{A_{Ecoli}}{V_{Ecoli}} \cdot h_i \cdot I$</td>
</tr>
<tr>
<td>(5.21) $M_{Y} \xrightarrow{k_{M_{Y}}} \emptyset$</td>
<td>$\alpha_{19} = \lambda_{M_{R}} \cdot M_R$</td>
</tr>
<tr>
<td>(5.22) $R \xrightarrow{k_{R}} \emptyset$</td>
<td>$\alpha_{20} = \lambda_{M_{Y}} \cdot M_Y$</td>
</tr>
<tr>
<td>(5.23) $R_2 \xrightarrow{k_{R}} \emptyset$</td>
<td>$\alpha_{21} = \lambda_{R} \cdot R$</td>
</tr>
<tr>
<td>(5.24) $Y \xrightarrow{k_{Y}} \emptyset$</td>
<td>$\alpha_{22} = \lambda_{R_2} \cdot R_2$</td>
</tr>
<tr>
<td>(5.25) $YI_{ex} \xrightarrow{k_{Y}} I$</td>
<td>$\alpha_{23} = \lambda_{Y} \cdot Y$</td>
</tr>
<tr>
<td>(5.26) $I_2R_2 \xrightarrow{k_{I}} 2I$</td>
<td>$\alpha_{24} = \lambda_{YI_{ex}} \cdot YI_{ex}$</td>
</tr>
</tbody>
</table>

$^{1}$ Variables without brackets denote number of molecules of the corresponding species.

$^{2}$ All propensity functions have units of min$^{-1}$.

$^{3}$ Avogadro's number: $N_A = 6.0221367 \times 10^{23}$ mmol$^{-1}$.

$^{4}$ For the stochastic modeling of $I_{ex}$ related processes see Appendix VI (page 178).
to the deterministic steady state
predictions as shown in Figure 5.5 (for
details about the simulation scheme
used, sampling method and the
calculation of statistics see Appendix V,
page 175). We observe that there is a
significant difference between the
average (stochastic) and the steady state
(deterministic) concentrations within a
wide range of extracellular IPTG concentrations \([I_{ex}]\). Additionally, as \([I_{ex}]\) increases, the
standard deviations become markedly high because of the positive feedback loop: the
total LacY concentration jumps between a maximal value corresponding to the induced
state and a minimal value corresponding to the uninduced state. This observation implies
that the marginal Probability Mass Function (PMF) for \([Y]_T\) contains important
information that cannot be captured with just the average and the standard deviation.
Furthermore, the marginal PMF for \([Y]_T\) is a sufficiently good observable since LacY is
the key species of the positive feedback mechanism. Thus, we do not need to consider the
entire 10-dimensional PMF for all the species.

Figure 5.6a depicts part of a sample path for the stochastically simulated \(lac\) operon
system and Figure 5.6b the estimated marginal PMF. A low extracellular IPTG
concentration has been used and the PMF appears unimodal with a heavy tail. On the
other hand, for a higher extracellular IPTG concentration (Figure 5.6c, d) the system
exhibits stochastically bistable behavior since the PMF is bimodal (Horstemke and
Figure 5.6: Typical sample path for the stochastically simulated lac operon system for $[I_{\text{ex}}] = 20 \mu\text{M}$ (panel a), and Probability Mass Function for this parameter set (panel b). For panels (c) and (d) $[I_{\text{ex}}] = 100 \mu\text{M}$. Parameters as in Table 5.2.

Lefever 1983). The existence of the heavy tail for low $[I_{\text{ex}}]$ and the bistability for higher $[I_{\text{ex}}]$ can be attributed to the autocatalytic effect from the action of the LacY permease resulting in the formation of two attracting vicinities. These two vicinities, one for low and one for high $[Y]_T$, are visited with higher probability. Thus, for low $[I_{\text{ex}}]$ the system “wanders” preferentially around the low attracting vicinity, but also spends some time around the high attracting vicinity, thereby generating the heavy tail observed in the PMF. For high $[I_{\text{ex}}]$, the system spends a significant amount of time around the high attracting vicinity, and thus a second mode in the PMF is created.

Bimodality of the PMF signals stochastic bistability; yet, one can also detect
bistability by analyzing the stationary potential, defined as (Gillespie 1979):

$$\varphi^s(x) = -\ln\left(\frac{P^s[X = x]}{P^s[X = 0]}\right) + c$$  \hspace{1cm} (5.39)

where \(c\) is an arbitrary constant and \(P^s[X = x]\) is the stationary probability that observable \(X\) will take the value \(x\). In our case \(X = [Y]_T\). To simplify our calculations we choose:

$$c = -\ln(P^s[X = 0]) \quad \Rightarrow \quad \varphi^s(x) = -\ln(P^s[X = x])$$  \hspace{1cm} (5.40)

The advantage of using \(\varphi^s\) for our purposes is that it "magnifies" minor "humps" of the PMF. Note that the maxima of the PMF appear as minima of the potential (attracting wells). Using this tool, we can analyze whether the stochastic model predicts the same bistable regime as the deterministic model. This comparison is of particular interest, since bistability is a key characteristic of the genetic architecture under consideration.

Hence, Figure 5.7a portrays the stationary potentials for \([Y]_T\) at various IPTG concentrations in comparison with the deterministic steady states for the nominal parameter set, which dictates slow \(lacY\) transcription and fast translation. The heavy tails of the PMF correspond to the extended plateaus in the potential. In Panel (b) the parameter set of Table 5.2 has been used, whereas the \(lacY\) transcription constants (\(k_{31MY}, k_{30MY}\)) have been taken to be 100 times higher and the translation constant (\(k_{4Y}\)) 100 times lower than the nominal values. Note that their product remains constant, and thus the deterministic steady states for the three parameter sets differ only in the \(lacY\) mRNA concentration (see equations AIV.6 and AIV.7). However, the stationary potentials for \([Y]_T\) are vastly different for these three parameter sets. We observe that the combination of fast transcription with slow translation, results in well defined modes in the PMF
Figure 5.7: The stationary potential for $[Y]_T$ compared to the deterministic bifurcation diagram for a range of IPTG concentrations: effect of stochasticity on the behavior of the system in different parameter regimes. Panel (a) nominal parameter set (Table 5.2). Panel (b): 100-fold fast $lacY$ transcription ($k_{sIMY} = 50 \text{ min}^{-1}$, $k_{sIMY} = 1 \text{ min}^{-1}$) and slow translation ($k_{sY} = 0.3 \text{ min}^{-1}$). Panel (c): 100-fold fast $lacI$ transcription ($k_{sMR} = 23 \text{ min}^{-1}$) and slow translation ($k_{sR} = 0.15 \text{ min}^{-1}$). Panel (d): simultaneous 100-fold fast $lacY$ and $lacI$ transcription and slow translation for both proteins ($k_{sIMY} = 50 \text{ min}^{-1}$, $k_{sIMY} = 1 \text{ min}^{-1}$, $k_{sMR} = 23 \text{ min}^{-1}$, $k_{sY} = 0.3 \text{ min}^{-1}$, $k_{sR} = 0.15 \text{ min}^{-1}$). Panel (e): 10-fold slower repressor-operator association and dissociation with only the 1st derepression mechanism, considering fast transcription and slow translation for $lacY$ ($k_e = 96 \text{ nM}^{-1} \text{min}^{-1}$, $k_{sR} = 0.24 \text{ min}^{-1}$, $k_{d2} = 0 \text{ nM}^{-2}$, $k_{d2} = 0 \text{ nM}^{-1} \text{min}^{-1}$, $k_{sIMY} = 50 \text{ min}^{-1}$, $k_{sIMY} = 1 \text{ min}^{-1}$, $k_{sY} = 0.3 \text{ min}^{-1}$). Panel (f): 10-fold faster repressor-operator association and dissociation with only the 1st derepression mechanism, considering fast $lacY$ transcription and slow translation ($k_e = 9600 \text{ nM}^{-1} \text{min}^{-1}$, $k_{sR} = 24 \text{ min}^{-1}$, $k_{d2} = 0 \text{ nM}^{-2} \text{min}^{-1}$, $k_{d2} = 0 \text{ nM}^{-1} \text{min}^{-1}$). Marked with $\bullet$ is the stochastic bistable (bimodal) regime and with $\star$ the deterministic. Surface color corresponds to stationary potential values and the bimodal regime is shaded green. Parameters as in Table 5.2 unless otherwise noted.
(heavy tails are no longer observed) and a sharper transition through the region of bistability. Furthermore, for this case the stochastically bistable regime is shifted closely to the deterministically bistable one.

These observations can be explained by the key role of LacY in the positive feedback mechanism: fast lacY transcription but slow translation means that the lacY mRNA dynamics will faithfully follow the current state of the operator. Noise in this case is produced at the translational level, but the operator fluctuations between the free and the repressor-bound state are still inherited to the LacY concentration and amplified through the positive feedback loop. However, when lacY is transcribed slowly, a small number of lacY mRNAs exist on average in the cell. Thus, it is “more difficult to detect” whether a change on lacY mRNA is due to stochasticity or due to a change in the current state of the repressor. This fuzziness is exhibited by the large plateaus in the potential and the destruction of bistability in comparison to the deterministic model’s predictions.

Such effects are not observed when the transcriptional and translational rates of only lacI are altered, as shown in panel (c), in which the lacI transcriptional rate (kₕMR) has been taken 100-fold higher and the translational rate (kₖR) 100-fold lower. This lack of effect is probably because LacI is not part of the positive feedback mechanism. Yet, if lacI and lacY transcriptional rates are simultaneously taken to be 100-fold higher (and their translational rates 100-fold lower, panel d), one observes potentials that are much sharper than in the case where only lacY transcription is fast (panel b). Moreover, the stochastically bistable regime seems to have shrunk in comparison to the deterministically bistable regime. These effects can be attributed to the negative feedback of LacI to the lacY transcription: when lacI transcription is slow (translation being fast)
the LacI concentration fluctuates stochastically within a large range of values, thereby
exerting negative feedback with a strength that varies stochastically. This variation
creates a blurring effect in the overall PMF. On the other hand, for fast lacI transcription
(slow translation), the fluctuations in the LacI concentration are narrowed, and thus the
strength of the negative feedback loop is less variable. As a result, the autocatalytic
mechanism can generate a sharp transition through the stochastically bistable regime, and
the modes of the PMF appear narrower and well defined.

The aforementioned simulations demonstrated the effect of transcriptional and
translational rates of lacI and lacY on the stationary potential (and the PMF). Yet, an
integral part of the lac operon system are the derepression mechanisms, the rate values of
which are expected to affect the stationary potential without necessarily affecting the
deterministic bifurcation structure. Thus, panels (e, f) of Figure 5.7 show the stationary
potentials for [Y]_T and the deterministic steady states versus IPTG concentrations in the
case where only the first derepression mechanism is taken into account (inducer binding
to free repressor, reactions 5.7, 8). For these simulations, fast lacY transcription and slow
translation was considered, in order to distinguish the effect of the derepression
mechanisms from the blurring and the plateaus generated from the low mRNA
concentrations.

In panel (e) the binding and unbinding rate constants are taken to be 10-fold lower
than those of the nominal case, whereas in panel (f) those rate constants are taken to be
10-fold higher than the nominal case (also, k_{dr2} and k_{-dr2} are both zero since only the first
derepression mechanism is functional). For both cases, the deterministic bifurcation
diagram is the same, since equation (5.31) at steady state gives:
\[
\frac{k_r}{k_{-r}} = \frac{[O]_r - [O]}{[R_2][O]}
\]

and thus if the binding affinity of the repressor to the operator \( k_r/k_{-r} \) is kept constant, the asymptotic solution of the deterministic system does not change. However, the stochastic bistable region is drastically affected and apparently it does not coincide with that of the deterministic model.

Specifically, Figure 5.7e demonstrates that for the stochastic model, slower repressor-operator interaction creates bimodality at an extended [IPTG] region before the left and after the right turning point. This phenomenon could be attributed to the dynamics of protein production being comparable to the repressor binding and unbinding to the operator. Therefore, the protein production machinery senses the stochastic fluctuations of the operator between the free and occupied state, and thus bistability is observed for a wide IPTG range. On the other hand, faster repressor-operator interaction reverts this extension of the bimodal regime. This reversion could be attributed to the time averaging that may occur for fast binding-unbinding dynamics: the protein production depends on the concentration of the free repressor. Thus, if the binding and unbinding is extremely fast, the protein production machinery will merely sense an average free repressor concentration for a wide range of external IPTG concentrations; consequently, bimodality will be suppressed.

Thus, we clearly see that biomolecular parameters not affecting the temporal asymptotic behavior of the deterministic model (such as the actual kinetic rates keeping their ratio constant) can have a strong effect on the behavior of the stochastic model. The region of the parameter space for which bistability is observed can be extended or shrunk.

Not all parameters were found to have such pronounced effects though: similar
simulations were performed for the case of the second derepression mechanism (inducer binding to repressor-operator complex, reactions 5.9, 10) by changing the rate values of the repressor-inducer-operator binding and unbinding ($k_{dr2}$, $k_{-dr2}$ of reactions 5.9, 10 to be 100-fold higher or 100-fold lower). However, no appreciable change of the bimodal regime was observed (results not shown). Furthermore, we observed only a slight extension of the bimodal regime when taking the repressor-inducer binding and unbinding rates ($k_{dr1}$, $k_{-dr1}$ of reactions 5.7, 8) to be 100-fold lower than the nominal values (results not shown). However, this effect was much weaker than that created by slower repressor-operator interaction. This lack of effect can be attributed to the fact that the aforementioned derepression and induction reactions (5.9, 10 and 5.7, 8 respectively) contain a species that exists in very high concentrations, namely the inducer IPTG. Thus, the stochasticity generated by those reactions is negligible in comparison to that generated by the repressor-operator binding and unbinding reactions, in which all participating species have low copy numbers. In view of this observation, we did not perform simulations with different rate constants for the IPTG facilitated import reactions (5.15 - 17), since they contain species with high copy numbers.

The PMF and the stationary potential presented so far contain information only at the asymptotic level, in the sense that they give probability of finding $[Y]_T$ at a specific value if we wait long enough. However, the stochastic model exhibits rich behavior also at the temporal level: in the bistable regime, noise-induced transitions between the two attracting vicinities are observed. The frequency of such transitions is quantified by the First Passage Time (FPT), which is the random time required for a transition from one attracting vicinity to another. The Mean First Passage Time (MFPT) is the mean of this
random time (van Kampen 1992). The MFPTs can be thought as quantifying the relative stability of the two attracting vicinities: the higher the MFPT for a transition from the lower to the upper vicinity, $\overline{T}_{\text{low} \rightarrow \text{up}}$, the more stable the lower vicinity, and the opposite. Further, the ratio of the MFPTs, $\overline{T}_{\text{low} \rightarrow \text{up}} / \overline{T}_{\text{up} \rightarrow \text{low}}$, is approximately equal to the ratio of the areas underneath the modes of the stationary PMF, $P^e_{\text{low}} / P^e_{\text{up}}$ (Gillespie 1981). In our simulations MFPTs and the standard deviations of FTPs were estimated from single-shot long simulations, typically $10^4$ - $10^6$ min, by using the vicinities $Y_T = 5$ molecules and $Y_T = 120$ molecules as reference low and high vicinities.

Figure 5.8 shows the MFPTs with error bars corresponding to the standard deviations of the First Passage Times for the case of fast lacY transcription and slow translation, in which the transition through the bistable regime is sharp. Circles are used to denote MFPTs for transitions from the lower to the upper attracting vicinities ($\overline{T}_{\text{low} \rightarrow \text{up}}$) and...
triangles for the opposite transitions ($\bar{T}_{\text{up\rightarrow low}}$). For low IPTG concentrations, the transitions from the lower to the upper vicinity are extremely rare, and if the system at some point approaches the upper vicinity state, it will rapidly transit towards the lower one. The opposite happens for high IPTG concentrations. Between these extreme cases there exists the bistable regime, in which a swapping of the relative magnitudes of the MFPTs $\bar{T}_{\text{low\rightarrow up}}$ and $\bar{T}_{\text{up\rightarrow low}}$ occurs.

A more detailed picture of the stochastic dynamics of transitions is given by the distribution of the FPTs (Figure 5.8b). In fact, it has been argued that for 1-dimensional stochastic systems the FPT follows approximately an exponential distribution (Procaccia and Ross 1977; Gillespie 1980), $p^s(\tau) = \lambda \cdot e^{-\lambda \tau}$. For multivariate systems with separation of time scales the FPT behaves approximately as in the one-dimensional case (Procaccia and Ross 1977). A statistical analysis of the FPTs for our lac operon model indeed shows an agreement between the observed distribution and the theoretically postulated exponential one, as Figure 5.8b reveals. This agreement also explains the fact that the estimated standard deviations are approximately equal to the value of the mean as shown in Figure 5.8a, since, for the exponential distribution, both are equal to $1/\lambda$.

We have so far discussed the behavior of the system in two levels: the first pertains to the temporal asymptotic behavior, which was analyzed in terms of the PMF and the stationary potential. The second level pertains to shorter time scales, those of the noise-induced transitions between the attracting vicinities, in the case where the system is stochastically bistable. We are now moving to a third level, in which it is also possible to analyze the much shorter time scales of the inherent noise created by the random occurrence of chemical reaction events. Our main tool for this investigation will be the
Figure 5.9: Panel (a): Autocovariance for extracellular IPTG $[I_{ex}] = 10 \mu M$. The autocovariance function decays exponentially fast, with an autocorrelation time $\tau_c \approx 25$ min for this case. Panel (b): dependence of the autocorrelation times on the induction levels as quantified by $[I_{ex}]$. Superimposed for comparison is the corresponding deterministic bifurcation diagram for $[Y]_T$, with arbitrarily scaled y-axis. Parameters as in Figure 5.7b.

\[
\kappa_X(t_1, t_2) := \left[(X(t_1) - \langle X(t_1) \rangle) \cdot (X(t_2) - \langle X(t_2) \rangle)\right]
\]

(5.42)

In our case $X = [Y]_T$. The autocovariance $\kappa_X(t_1, t_2)$ quantifies trends between two points in time, and, for a wide sense stationary process it is only a function of the time lag; thus, $\kappa_X(t, t+\tau) = \kappa_X(\tau)$. Therefore, this parameter shows the time window for which the system retains its memory. Once we have computed the autocovariance as predicted by our model, we can compare with experimental data (Rosenfeld et al. 2005).

Figure 5.9a shows that $\kappa_X(\tau)$ decays exponentially fast as $\tau$ increases. Thus, the system progressively "forgets" its previous history. The autocorrelation time $\tau_c$ is indicative of the time scales for which the system still "remembers" its history, and was computed from the equation:

\[
\frac{\kappa_X(\tau_c)}{\kappa_X(0)} = \frac{1}{e}
\]

(5.43)

In Figure 5.9b the autocorrelation time ($\tau_c$) is plotted with respect to the extracellular
IPTG concentration. Clearly, \( \tau_c \) depends on the IPTG concentration and, for the bistable regime, it assumes values significantly larger than those pertaining to the monostable regimes (50 min in the bistable versus < 10 min in the monostable). This effect can be attributed to the noise-induced transitions that follow much slower dynamics than the intrinsic noise, and thus result in much higher autocorrelation times in the stochastic bistable regime. Indeed, the maximum autocorrelation time \( (\tau_{c,\text{max}} \approx 50 \text{ min}) \) occurs for \( [I_{ex}] \approx 20 \mu M \) which is outside of the deterministic bistable region, but it corresponds to the point where the system spends approximately the same mean time in the upper and lower attracting vicinity. The latter fact is revealed by observing Figure 5.8a: the graphs for \( \bar{T}_{\text{low} \rightarrow \text{up}} \) and \( \bar{T}_{\text{up} \rightarrow \text{low}} \) versus \( [I_{ex}] \) cross at \( [I_{ex}] \approx 20 \mu M \); therefore, at this point the noise-induced transitions have a maximal contribution in the value of \( \tau_c \). Yet, since these transitions follow much slower dynamics than the intrinsic noise, the autocorrelation time is maximized therein. In view of these observations, the assertion by Rosenfeld et al. (2005) that intrinsic noise has smaller autocorrelation times than extrinsic noise may not be true for a class of genetic architectures and induction conditions.

5.5 Conclusion

Starting from a network of reactions that incorporates biological information for the \( lac \) operon genetic switch, we have derived a deterministic and the corresponding stochastic model for this system.

According to both models, the system exhibits bistability. The deterministic model predicts that certain biomolecular parameters can greatly affect the extent and positioning of the bistable regime. Specifically, stronger repressor-operator binding, weaker
repressor-inducer binding, or lower total operator concentrations shift the bistable regime to higher extracellular IPTG concentrations. On the other hand, there are parameter changes that do not affect the steady state $[Y]_T$ concentration of the deterministic model, but strongly affect the stationary behavior of the stochastic model. Thus, such effects are genuine outcomes of stochasticity. Specifically, faster $lacY$ transcription and slower translation rates result in sharper transition through the stochastically bistable regime and narrower probability distributions. These observations may be explained by the noise generated at the transcriptional level being smaller in the case of fast $lacY$ transcription. Thus, the operator fluctuations can be faithfully amplified through the positive feedback loop and produce a clearly bistable behavior.

Moreover, slower repressor-operator binding and unbinding rates, result in a broadening of the stochastically bistable region, even if the repressor-operator dissociation constant remains the same. Such effects could be attributed to the operator fluctuations being “inherited” to the LacY time course in the case when both the LacY and $lacO$ dynamics have similar time scales. Furthermore, faster LacY production and degradation dynamics have a similar effect in broadening the stochastic bistable regime. Since bistability is a key characteristic of the behavior of the $lac$ operon system, it follows that stochasticity can enhance the robustness of the system by extending the bistable region.

The relative stability of the attracting vicinities generated in the stochastic model can be quantified by the First Passage Times, namely the random times needed for a jump between the attracting vicinities. The First Passage Times follow approximately exponential distributions, and their means were found to depend strongly on the
extracellular IPTG concentration. For higher inducer concentrations the upper attracting vicinity is more stable, resulting in lower Mean First Passage Times for an upwards jump (from the lower to the upper vicinity). Stability is interchanged for lower IPTG concentrations.

Moreover, the stochastic model makes it possible to analyze the statistics of the random fluctuations in the concentrations of LacY at stationary phase. The autocovariance function was found to decay exponentially with autocorrelation times highly dependent on the induction level as quantified by the extracellular IPTG concentration. Therefore, in the monostable regimes these times were significantly smaller than those observed in the bistable regime, an effect that can be attributed to the noise-induced transitions.

In conclusion, we have presented a comparative analysis of corresponding deterministic and stochastic models for the lac operon system. The incorporation of biological information into the models revealed the effect of biomolecular parameters in presence and in absence of stochasticity. These parameters can be modulated experimentally, and the presented models can provide valuable insight into the underlying biological processes. More importantly, the comparison of the predictions of the two models made possible the isolation of the genuine effects of stochasticity on key characteristics of the lac operon behavior, such as the extent of the bistable regime and the robustness of the system.
Chapter 6:

Quantifying Cell Population Heterogeneity for the \textit{lac} operon

Genetic Network

In this chapter, we will assess the effect of the major sources of stochasticity discussed in Chapter 2 on the cell population heterogeneity for the \textit{lac} operon genetic network. To this end, we will incorporate the reaction network developed in Chapter 5 in the cell population algorithm developed in Chapter 3, considering two cases for the dynamics of intracellular reactions: first, that reactions are deterministic, and second, that they are stochastic. We will subsequently perform simulations to elucidate how cell population heterogeneity is shaped by the two sources of noise, namely the intrinsic, which is stochasticity in intracellular reactions, and the extrinsic, which are stochastic DNA duplication, division and partitioning.

6.1 Deterministic Reaction Dynamics

As a start for our analysis, we will assume that reaction dynamics are deterministic inside the cell. This assumption of course, is not true, due to the small number of molecules that interact thereby giving rise to intrinsic stochasticity in the system.
However, we would first like to focus on the effects of extrinsic noise sources and then incorporate intrinsic stochasticity to quantify its effect.

6.1.1 Cell Population M-Equation with Deterministic Reaction Rates

In Chapter 3 we discussed the formulation of a M-Equation that pertains to cell populations and treats reaction kinetics as stochastic processes and cell growth as a deterministic process. Thus, the extension of this approach to the case where reactions are deterministic is straightforward: reactions will now be treated in the same way as the cell growth as shown in the following equation:

$$
\frac{\partial}{\partial t} J_v((X_1, V_1),\ldots,(X_i, V_i),\ldots,(X_v, V_v); t) = \nabla_{X_i} \left[ r(X_i, V_i) \cdot J_v((X_1, V_1),\ldots,(X_i, V_i),\ldots,(X_v, V_v); t) \right] - \sum_{\zeta=1}^{v} a_s (X_{\zeta} - v_s, V_{\zeta}) \cdot J_v((X_1, V_1),\ldots,(X_{\zeta} - v_s, V_{\zeta}),\ldots,(X_v, V_v); t) - a_s (X_{\zeta}, V_{\zeta}) \cdot J_v((X_1, V_1),\ldots,(X_{\zeta}, V_{\zeta}),\ldots,(X_v, V_v); t)$$

$$- \sum_{\zeta=1}^{v} \frac{\partial}{\partial V_{\zeta}} \left[ g(X_{\zeta}, V_{\zeta}) \cdot J_v((X_1, V_1),\ldots,(X_{\zeta}, V_{\zeta}),\ldots,(X_v, V_v); t) \right] + 2 \sum_{\zeta=1}^{v-1} \sum_{\zeta'=\zeta+1}^{v} a_d (X_{\zeta} + X_{\zeta'}, V_{\zeta} + V_{\zeta'}) \cdot h(X, V|X_{\zeta} + X_{\zeta'}, V_{\zeta} + V_{\zeta'}) \cdot J_{v-1}((X_1, V_1),\ldots,(X_{\zeta} + X_{\zeta'}, V_{\zeta} + V_{\zeta'}),\ldots,(X_{\zeta-1}, V_{\zeta-1}),\ldots,(X_{\zeta+1}, V_{\zeta+1}),\ldots,(X_v, V_v); t) - \sum_{\zeta=1}^{v} a_d (X_{\zeta}, V_{\zeta}) \cdot J_v((X_1, V_1),\ldots,(X_{\zeta}, V_{\zeta}),\ldots,(X_v, V_v); t)$$

(6.1)

Note the following differences between the above equation (6.1) and equation (3.15):

first, the reaction terms here (namely in eq. 6.1) have the differential form highlighted in the red box.

Furthermore, for the partitioning function h, the components that express species
partitioning can no longer be given by equations (2.32) and (2.37), since now the species to be partitioned are not discrete entities. Thus, new expressions have to be formulated:

\[ c_i \left( X_{d,i}^{\text{DNA},i} \mid X_{m}, V_{m}, V_d \right) = \prod_{j=1}^{n} \delta \left( X_{d,j}^{\text{DNA},i} - \frac{1}{2} X_{m,j}^{\text{DNA},i} \right) \quad \text{for } i = 1, ..., d \]  

\[ b_i \left( X_{d,i} \mid X_{m,i}, V_m, V_d \right) = \delta \left( X_{d,i} - \frac{V_d}{V_m} \cdot X_{m,i} \right) \quad \text{for } i = 1, ..., n \]  

Hence, the content for non-chromosomal DNA species is partitioned between the two daughter cells proportionally to each daughter’s volume. For chromosomal DNA species equal partitioning is always occurring. Intuitively, limiting agreement between equations (6.2, 3) and (2.32, 37) respectively is guaranteed by the behavior of the binomial and the hypergeometric distributions, which approach Dirac delta functions at the limit of infinitely many trials. Partitioning of the volume is still taken to follow a symmetric beta:

\[ \beta \left( V_d \mid V_m \right) = \frac{1}{V_m^q \cdot \Gamma (2 \cdot q) \cdot \left( \frac{V_d}{V_m} \right)^{q-1}} \left( 1 - \frac{V_d}{V_m} \right)^{q-1} \]  

and the full partitioning function is given as previously:

\[ h \left( X_d, V_d \mid X_m, V_m \right) = \beta \left( V_d \mid V_m \right) \cdot \prod_{i=1}^{n} b_i \left( X_{d,i} \mid X_{m,i}, V_m, V_d \right) \cdot \prod_{i=1}^{d} c_i \left( X_{d,i}^{\text{DNA},i} \mid X_{m,i}, V_m, V_d \right) \]  

Finally, the species contents \( X_i \) are now continuous variables that change due to reactions as prescribed by the system of differential equations (for notation refer to paragraph 2.2.1, Chemical Reactions):

\[ \frac{dX_i}{dt} = r_i \left( X, V \right) = \sum_{j=1}^{m} \left( \beta_{ij} - \alpha_{ij} \right) \cdot k_j \cdot V \cdot \prod_{i=1}^{N} \left( \frac{X_i}{V} \right)^{\alpha_{ij}} \quad \text{for } i = 1, ..., N \]  

\[ \frac{dV}{dt} = g \left( X, V \right) \]
Therefore, the net outflow of probability from state \((v, (X_1, V_1), ..., (X_\zeta, V_\zeta), ..., (X_\nu, V_\nu))\)
due to reactions occurring only in cell \(\zeta\) will be given as:

\[-\nabla_{X_\zeta} \cdot \left[ r(X_\zeta, V_\zeta) \cdot J(v((X_1, V_1), ..., (X_\zeta, V_\zeta), ..., (X_\nu, V_\nu); t)) \right] \]  
(6.7)

where \(r\) is the vector with the rate expressions for the production of species \(X\) (with the
convention that \(r_i\) is positive when \(X_i\) is produced and negative when \(X_i\) is depleted), and
the divergence \(\nabla_{X_i}\) operates in a vector function \(Z\) as:

\[\nabla_{X_i} \cdot Z = \sum_{i=1}^{N} \frac{\partial Z_i}{\partial X_{i,j}}\]  
(6.8)

For the purpose of simulating the \(lac\) operon genetic network, the rates \(r_i\) will be given
by the extensive form of equations (5.28 - 36), which is shown below (species without
brackets denote content in equations (5.28 - 36), which is shown below (species without
brackets denote content in molar units):

\[\frac{dM_{R}}{dt} = k_{sR} \cdot V - \lambda_{MR} \cdot M_{R}\]  
(6.9)

\[\frac{dR}{dt} = k_{sR} \cdot M_{R} - 2 \cdot \frac{k_{R}}{V} \cdot R^2 + 2 \cdot k_{-R} \cdot R - \lambda_{R} \cdot R\]  
(6.10)

\[\frac{dR_2}{dt} = \frac{k_{R} \cdot R^2}{V} - k_{-R} \cdot R - \frac{k_{R}}{V} \cdot R_{2} \cdot O + k_{-R} \cdot R_{2} \cdot O - \frac{k_{dr1}}{V^2} \cdot R_{2} \cdot I^2 + k_{-dr1} \cdot I_2 R_2 - \lambda_{R} \cdot R_2\]  
(6.11)

\[\frac{dO}{dt} = - \frac{k_{R} \cdot R_{2} \cdot O + k_{-R} \cdot R_2 O + \frac{k_{dr2}}{V^2} \cdot R_2 O \cdot I^2 - \frac{k_{dr2}}{V} \cdot O \cdot I_2 R_2}{V}\]  
(6.12)

\[\frac{dR_2 \cdot O}{dt} = \frac{k_{R} \cdot R_{2} \cdot O - k_{-R} \cdot R_2 O - \frac{k_{dr2}}{V^2} \cdot R_2 O \cdot I^2 + \frac{k_{dr2}}{V} \cdot O \cdot I_2 R_2}{V}\]  
(6.13)

\[\frac{dI}{dt} = -2 \cdot \frac{k_{dr1}}{V^2} \cdot R_{2} \cdot I^2 + 2 \cdot k_{-dr1} \cdot I_2 R_2 - 2 \cdot \frac{k_{dr2}}{V^2} \cdot R_2 O \cdot I^2 + 2 \cdot \frac{k_{dr2}}{V} \cdot O \cdot I_2 R_2 + k_{R} \cdot Y_{ex} + h_{I} \cdot A \cdot [I_{ex}] - h_{I} \cdot A \cdot \frac{A}{V} \cdot I + 2 \cdot \lambda_{I2R2} \cdot I_2 R_2 + \lambda_{Y_{ex}} \cdot Y_{ex}\]  
(6.14)
\[
\frac{dI_2R_2}{dt} = \frac{k_{dR1}}{V^2} \cdot R_2 \cdot I^2 - k_{dR2} \cdot I_2R_2 + \frac{k_{dO2}}{V} \cdot R_2 \cdot O \cdot I^2 - \frac{k_{dR2}}{V} \cdot O \cdot I_2R_2 - \lambda_{12R2} \cdot I_2R_2 \tag{6.15}
\]

\[
\frac{dM_Y}{dt} = k_{s0MY} \cdot R_2O + k_{s1MY} \cdot O - \lambda_{MY} \cdot M_Y \tag{6.16}
\]

\[
\frac{dY}{dt} = k_{sY} \cdot M_Y + \left( k_{fr} + k_{fr_p} \right) \cdot YI_{ex} - k_p \cdot \left[ I_{ex} \right] \cdot Y - \lambda_Y \cdot Y \tag{6.17}
\]

\[
\frac{dYI_{ex}}{dt} = - \left( k_{fr} + k_{fr_p} \right) \cdot YI_{ex} + k_p \cdot \left[ I_{ex} \right] \cdot Y - \lambda_{YI_{ex}} \cdot YI_{ex} \tag{6.18}
\]

The functions \( a_s, g, \) and \( a_d \) are given by equations (2.29, 27), and (2.30). Simulation of the M-Equation (6.1) is performed with the algorithm discussed in Chapter 3, but without the occurrence of discrete chemical reaction events; rather, continuous reaction kinetics are simulated between the DNA duplication and division events.

### 6.1.2 Simulation of Phenotypic Distributions and Statistics

In order to study the population behavior, MC simulations of the M-Equation (6.1) were performed with the kinetic constants shown in Table 5.2 (note though that the volume, area and total operator content are now functions of time) and for the parameter values given in Table 6.1.

Two representative transient simulations for all the cells in the population as well as for the population average and standard deviations for \( \text{Lac}Y_T \) are shown in Figure 6.1.

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Units</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>( g )</td>
<td>( \text{min}^{-1} )</td>
<td>0.0231</td>
</tr>
<tr>
<td>( n_d )</td>
<td>( \text{dim/less} )</td>
<td>25</td>
</tr>
<tr>
<td>( V_{d,\text{crit}} )</td>
<td>( \text{L} )</td>
<td>( 11 \cdot 10^{-16} )</td>
</tr>
<tr>
<td>( q )</td>
<td>( \text{dim/less} )</td>
<td>80</td>
</tr>
<tr>
<td>( n_g )</td>
<td>( \text{dim/less} )</td>
<td>25</td>
</tr>
<tr>
<td>( V_{s,\text{crit}} )</td>
<td>( \text{L} )</td>
<td>( 8 \cdot 10^{-16} )</td>
</tr>
</tbody>
</table>
Figure 6.1: Panel (a): Transients for all the cells in the population starting from a single cell close to the lower state. Panel (b): Dynamics of the population mean and standard deviation for the total LacY concentration. Panels (c, d): As in panels (a, b) starting with a cell close to the upper state. $[I_{\text{tot}}] = 22 \, \mu\text{M}$ and all other parameters as in Table 5.2 and Table 6.1.

Both simulations (panels a, b and c, d) are initialized with a single cell, which gives rise to about 7000 - 8000 cells in the end of the simulation (400 min). For panels (a, b) the cell is close to the lower steady state predicted by the deterministic single cell model, whereas for panels (c, d) the initial cell is close to the upper steady state.

The simulations show that the cells do not switch between the two states and the dynamics of the population average exhibit oscillations whose amplitude decrease with time. However, due to computational time limitations, we could not simulate the cell population up to the point that the average reaches a steady value. To overcome these limitations we employed a Constant Number MC scheme, in which a maximum number
Figure 6.2: Panel (a): Dynamics of the population average for the total LacY concentration. The simulation was initialized with a single cell close to the lower attracting state. The black curve shows the number of cells in the population and the vertical dashed red line shows the time when the Constant Number MC Scheme was triggered. The maximum allowed number of cells in the population was $10^4$. Panel (b): As in panel (a) with the initial cell close to the upper attracting state. Panels (c, d): Timecourses for single cells: only one daughter is followed after division, and thus a line of cells is simulated, see also Figure 2.1. The cell was initialized close to the lower (c) or upper (d) attracting state. No transitions between attracting states are observed even for large times. $[I_{ex}] = 22\, \mu$M and all other parameters as in Table 5.2 and Table 6.1.

of observed cells, say $N_{\text{cells max}} = 2000$ is defined. If, after a division event, the population size surpasses the maximum (having 2001 cells) then the algorithm removes randomly one cell from the population to restore the population size to the maximum allowed number of cells (Smith and Matsoukas 1998; Lee and Matsoukas 2000). Using this technique, we were able to extend the simulated time horizon, up to the point where the population averages reach a constant value (Figure 6.2a, b). Furthermore, it is important to note that cell-line simulations ($N_{\text{cells max}} = 1$, following only one daughter at
each division) showed that even for large simulated times (∼ 6000 min ≈ 4 days), no transitions between the upper and the lower attracting state were observed (Figure 6.2c, d). This result does not necessarily mean that transitions are impossible in this model, but that they are extremely rare.

From the transient simulations of the cell population just discussed, one can obtain phenotypic distributions for the total LacY concentration $[Y]_T$ (Figure 6.3). Such distributions are shown for three different extracellular IPTG concentrations in panels (a, d, g). In order to make the connection with experiments possible, we also plot the distributions of the log($[T]_T$) in panels (b, e, h). Such distributions can be obtained by incorporating a fluorescent protein reporter into the genetic network and measure single cell fluorescence (FL1-channel) with FC using logarithmic amplification. Note that these logarithmic plots, as they correspond to the Number Density Function (NDF) of log($[Y]_T$), may distort the distributions, such that heavy tails are mistaken for distinct upper modes. Thus, caution must be exercised when one tries to identify parameter regions where the system is bistable: one has to always refer to the density of the original quantity (in our case $[Y]_T$), and not its logarithm.

Another way to visualize data obtained from simulations is by utilizing scatter plots for the size versus the total LacY content $Y_T$. Such scatter plots can also be obtained with FC by measuring and plotting Forward Scatter (FSC-channel) versus Fluorescence (FL1-channel). In these plots, each point corresponds to one cell and the color represents the density of points per unit area of the plane. Distinct subpopulations of cells can be easily identified as separate clouds of points in the graph, similarly to the existence of distinct modes in the distributions.
Figure 6.3: Panel (a): NDF for the total LacY concentration for $[I_{ex}] = 10$ µM at time $t = 446.3$ min where the number of cells was $N = 47826$. Panel (b): As in panel (a) but for the logarithm of the total LacY concentration (this is what a distribution obtained by FC with logarithmic amplification would look like). Panel (c): Scatter plot for the volume of the cells versus the total LacY content corresponding to panel (a) (this is what a FSC-FL1 scatter plot obtained by FC with logarithmic amplification in both channels would look like). Panels (d, e, f): As in panels (a, b, c) with $[I_{ex}] = 20$ µM, at time $t = 446.1$ min where the number of cells was $N = 47595$. Panels (g, h, i): As in panels (a, b, c) with $[I_{ex}] = 50$ µM, at time $t = 446.1$ min where the number of cells was $N = 47611$. All other parameters as in Table 5.2 and Table 6.1.

Thus, both the $[Y]\_T$ distributions and the scatter plots show that for the intermediate IPTG concentration, $[I_{ex}] = 20$ µM, the population exhibits bistability: the distributions are bimodal and two subpopulations are observed in the scatter plot. Note, however, that this simulation was initiated with two cells, one close to the upper and one close to the lower state. Since the transients discussed earlier (Figure 6.2a, b) show that no transitions
between the two attracting states are observed, one can experimentally observe only hysteresis: an overnight culture grown with high IPTG will equilibrate close to the high state upon inoculation to an intermediate IPTG concentration. Similarly, an overnight culture grown in absence of IPTG will equilibrate close to the low mode.

It is further interesting to observe that the cell subpopulations in the scatter plots appear to be S-shaped. One possible explanation for this phenomenon is that the DNA duplication, which happens roughly in the middle of the division cycle, results in higher RNA, and thus protein production rates. Due to the autocatalytic nature of the lac operon dynamics, the LacY production rate is amplified even more, an effect that shapes the upper arm of the cloud of points.

6.1.3 Comparison of an Average Single Cell Model with the Cell Population Model

As discussed in the previous section, simulating in detail the intra- and inter-cellular processes at the population level is computationally demanding. Therefore, it is natural to ask whether one can adequately predict the dynamics of the population average with the use of a single cell model that apparently neglects heterogeneity.

Preliminary studies by Mantzaris (2004; 2005a; 2006; 2007) showed marked deviations between the single cell model behavior and the cell population behavior, thereby giving a negative answer to this question. In particular, Manztaris (2004; 2005a; 2006; 2007) compared the steady state behavior of continuum models with the stationary average of the “corresponding” CPBs. He thus showed that the regions in which the two models exhibit specific types of behavior such as bistability are very different. Mantzaris interpreted these differences as genuine effects of cell population heterogeneity.

In these studies, the single cell models that are incorporated to the CPB are written for
species concentrations. However, the CPB accepts single cell models written for cellular contents and not concentrations (Fredrickson 1976). The concentrations are intensive variables in contrast to the contents which are extensive; one immediate consequence of this fact, is that it is incorrect to partition the concentrations upon cell division.

Moreover, the cell volume is not taken into account in any of these studies. However, cells in the exponential phase are continuously growing and dividing, thereby changing their volume roughly two-fold during one division cycle. In turn, the change in volume shifts the thermodynamic equilibrium point and also affects the dynamics if multi-molecular reactions are present in the reaction scheme, which is the case in all of the systems studied by Mantzaris. Therefore, neglecting cell volume creates a priori an inconsistency between the single cell expressions incorporated into the population balance and the single cell continuum model that neglects heterogeneity.

Furthermore, in some of these studies unequal partitioning has been used to essentially artificially generate complex behaviors, such as oscillations in a reaction network with 0th and 1st order reactions. However, the asymmetry in E. coli division is negligible: cells may stochastically divide in two unequal daughters but consistent generation of one large and one small daughter has not been experimentally observed. In fact it has been shown that the distribution of daughter cell sizes has a mean corresponding to equal partitioning and a small coefficient of variation (Marr et al. 1966; Harvey et al. 1967).

Finally, DNA duplication was never taken into account in any of the studies by Mantzaris (2004; 2005a; 2006; 2007). In contrast, the total operator content (or concentration) has been assumed to be constant (Mantzaris 2006; 2007). However, DNA duplication results in different metabolic fluxes during one division cycle, and thus needs
to be incorporated as another source of heterogeneity.

In view of these observations we need to reexamine the question of adequately predicting the population average dynamics with the use of a single cell model. Our cell population model does not suffer from any of the dubious modeling practices just discussed. The single cell model incorporated into our population simulations pertains to species contents and not concentrations, cell volume is explicitly taken into account, and DNA duplication is modeled as discrete event generating free operator content. We further need to carefully construct an average single cell model for the comparison to be fair.

This average single cell model, essentially consists of the deterministic single cell equations (5.28 - 36), with the dilution due to growth taken into account. Since we have assumed exponential growth, the average specific growth rate is:

\[
\frac{1}{V} \frac{dV}{dt} = g
\]

(6.19)

and therefore the average single cell model is written as follows:

\[
\frac{d[M_R]}{dt} = k_{sMR} \cdot [M_R] - \lambda_{MR} \cdot [M_R] - g \cdot [M_R]
\]

(6.20)

\[
\frac{d[R]}{dt} = k_{dr} \cdot [M_R] - 2 \cdot k_{2R} \cdot [R]^2 + 2 \cdot k_{-2R} \cdot [R_2] - \lambda_R \cdot [R] - g \cdot [R]
\]

(6.21)

\[
\frac{d[R_2]}{dt} = 2k_{R} \cdot [R]^2 - k_{-2R} \cdot [R_2] - k_{r} \cdot [R_2] \cdot [O] + k_{r} \cdot \left( ([O]_R - [O]) - k_{dr1} \cdot [R_2] \cdot [I]^2 + k_{-dr1} \cdot [I_2R_2] - \lambda_{R2} \cdot [R_2] - g \cdot [R_2]
\]

(6.22)

\[
\frac{d[O]}{dt} = -k_{r} \cdot [R_2] \cdot [O] + k_{-r} \cdot \left( ([O]_R - [O]) + k_{dR_2} \cdot \left( ([O]_R - [O]) \cdot [I]^2 + k_{-dR_2} \cdot [O] \cdot [I_2R_2]
\]

(6.23)
\[
\frac{d[I]}{dt} = -2 \cdot k_{d1} \cdot [R_2] \cdot [I]^2 + 2 \cdot k_{\text{d1}} \cdot [I_2 \cdot R_2] - 2 \cdot k_{\text{d2}} \cdot \left( \frac{[O]}{V} - [O] \right) \cdot [I]^2 \\
+ 2 \cdot k_{\text{d2}} \cdot [O] \cdot [I_2 \cdot R_2] + k_{Y} \cdot [YI_{\text{ex}}] + h_{t} \cdot \left( \frac{A}{V} \right) \cdot \left( [I_{\text{ex}}] - [I] \right) \\
+ 2 \cdot \lambda_{2R_2} \cdot [I_2 \cdot R_2] + \lambda_{YI_{\text{ex}}} \cdot [YI_{\text{ex}}] - g \cdot [I] 
\] (6.24)

\[
\frac{d[I_2 \cdot R_2]}{dt} = k_{\text{d1}} \cdot [R_2] \cdot [I]^2 - k_{\text{d1}} \cdot [I_2 \cdot R_2] + k_{\text{d2}} \cdot \left( \frac{[I]}{V} - [I] \right) \cdot [I]^2 \\
- k_{\text{d2}} \cdot [O] \cdot [I_2 \cdot R_2] - \lambda_{2R_2} \cdot [I_2 \cdot R_2] - g \cdot [I_2 \cdot R_2] 
\] (6.25)

\[
\frac{d[M_{V}]}{dt} = k_{sO_{MY}} \cdot \left( \frac{[O]}{V} - [O] \right) + k_{sIMY} \cdot [O] - \lambda_{MY} \cdot [M_{V}] - g \cdot [M_{V}] 
\] (6.26)

\[
\frac{d[Y]}{dt} = k_{Y} \cdot [M_{V}] + (k_{n} + k_{p}) \cdot [YI_{\text{ex}}] - k_{p} \cdot [Y] \cdot [I_{\text{ex}}] - \lambda_{Y} \cdot [Y] - g \cdot [Y] 
\] (6.27)

\[
\frac{d[YI_{\text{ex}}]}{dt} = -(k_{n} + k_{p}) \cdot [YI_{\text{ex}}] + k_{p} \cdot [Y] \cdot [I_{\text{ex}}] - \lambda_{YI_{\text{ex}}} \cdot [YI_{\text{ex}}] - g \cdot [YI_{\text{ex}}] 
\] (6.28)

Note that we did not include the dilution effect for the operator species O since this species is continuously regenerated by the DNA duplication process. Also, notice that we need an estimate for the average membrane area over the volume, \( \langle A/V \rangle \) and the average total operator concentration \( \langle [O_{I}] \rangle \). The former quantity could be estimated as follows: assume that a newborn cell with volume \( \frac{1}{2} V_{d, \text{crit}} \) divides when it reaches \( V_{d, \text{crit}} \). Then, since the cell is growing exponentially and the \( A/V \) is known from equation (2.25):

\[
\frac{A}{V} = \frac{10}{13} \cdot \frac{1}{13} \cdot \frac{1}{3} \cdot \frac{2}{3} \cdot \pi \cdot \frac{1}{\sqrt{3}} 
\] (6.29)

The average \( \langle A/V \rangle \) can be calculated over the timecourse of a birth-division cycle, in accordance with the first mean value theorem for integration:
\[
\langle \frac{A}{V} \rangle = \frac{1}{t_{\text{division}} - t_{\text{birth}}} \int_{t_{\text{birth}}}^{t_{\text{division}}} \frac{A(t)}{V(t)} \, dt
\]
\[
= \frac{10}{13} \cdot \frac{g}{\ln(2)} \cdot \frac{1}{13^3} \cdot \frac{2}{3} \cdot \frac{1}{\pi^3} \cdot \int_{0}^{\frac{\ln(2)}{g}} \left( \frac{1}{2} \cdot V_{\text{d, crit}} \cdot e^{\varepsilon t} \right)^{\frac{1}{3}} \, dt
\]

(6.30)

Using the parameters of Table 6.1, this estimated average evaluates to \(6.0 \cdot 10^5\) dm\(^{-1}\). Cell population simulations (Figure 6.4a) give an average equal to \(6.2 \cdot 10^5\) dm\(^{-1}\). Thus, using these heuristic arguments, we estimated the average ratio \(\langle A/V \rangle\) with a remarkably low error (3.2%).

For estimating the latter quantity \(\langle [O]_T \rangle\), let us assume that a newborn cell with volume \(\frac{1}{2}V_{\text{d, crit}}\) duplicates its DNA when it reaches \(V_{s, \text{crit}}\) and of course divides when it reaches \(V_{\text{d, crit}}\). Then, since the cell is growing exponentially and the initial operator content is 1 copy before duplication:

\[
\langle [O]_T \rangle = \frac{1}{t_{\text{division}} - t_{\text{birth}}} \left( \int_{t_{\text{birth}}}^{t_{\text{duplication}}} \frac{1}{V(t)} \, dt + \int_{t_{\text{duplication}}}^{t_{\text{division}}} \frac{2}{V(t)} \, dt \right)
\]
\[
= \frac{g}{\ln(2)} \cdot \left( \int_{0}^{\frac{1}{2} \cdot V_{\text{d, crit}}} \frac{1}{N_A \cdot \frac{1}{2} \cdot V_{\text{d, crit}} \cdot e^{\varepsilon t}} \, dt + \int_{\frac{1}{2} \cdot V_{\text{d, crit}}}^{\frac{1}{2} \cdot V_{\text{d, crit}}} \frac{2}{N_A \cdot \frac{1}{2} \cdot V_{s, \text{crit}} \cdot e^{\varepsilon t}} \, dt \right)
\]

(6.31)

Using the parameters of Table 6.1, this estimated average evaluates to 3.0 nM. Cell population simulations (Figure 6.4b) give an average equal to 3.1 nM (error 3.2%).

Finally, the results of the comparison of the average single cell model (6.20 - 28) with the cell population model (6.1 - 18) appear in Figure 6.4c. The blue and red curves show bifurcation diagrams for the total steady state LacY concentration calculated by the single cell model. For the blue curve, the average values for the membrane area over cell volume ratio and the total operator concentration were taken to be as shown in panels (a)
Figure 6.4: Panels (a, b): Dynamics of the population mean for the membrane area over the volume ratio and the total operator concentration. Panel (c): agreement between the cell population model and the single cell model when the average values for $A/V$ and $[O_T]$ are used. For the blue curve, the average values for $A/V$ and $[O_T]$ were taken to be as shown in panels (a) and (b). For the red curve, the average values were taken to be equal to the estimates of equations (6.30, 31). Panel (d): a representative bimodal NDF and the corresponding averages and standard deviations for $[I_{ex}] = 20 \ \mu M$. Parameters as in Table 5.2 and Table 6.1.

and (b). For the red curve, these average values were taken to be equal to the estimates of equations (6.30, 31). The error bars correspond to the predictions of the cell population model for the average and standard deviation of the modes of the resulting NDFs (each segment of the error bar is one standard deviation, so the full line shows two standard deviations, see also panel d).

The qualitative agreement between the single cell steady states and the population averages is excellent. From a quantitative perspective, the agreement of the population model with the single cell average model is also good when the population average
values for are used for A/V and \([O_T]\) in the single cell model. In the case where the estimated values from equations (6.30, 31) are used, the total LacY concentration at maximal induction is slightly underestimated due to the underestimation of the average total operator concentration. Thus, it turns out that the level of agreement between the steady states of the single cell model and the averages of the population model, depends on how good are the estimates for the average cell characteristics.

This agreement between the single cell and the cell population model can be explained as follows. In this simple system, the only coupling between the cells is due to stochastic partitioning which can generate variability but no bias on the total LacY concentration, in the sense that one does not observe consistently higher LacY concentrations in one daughter versus the other. In fact, the two newborn daughters may have different LacY_T contents, but they always have equal LacY_T concentrations, which are identical to their mother's LacY_T concentration just before division. Similarly, unsynchronized DNA duplication and division events also contribute to the observed variability in LacY_T, but they cannot consistently bias the LacY concentration.

In general, since the population dynamics emerge from the single cell dynamics, unless there is either a strong intercellular coupling mechanism, or differential growth, division, or death rates for distinct phenotypes, there is no reason to expect significant deviations between the predictions of the single cell and the cell population models. In our system, clearly such complex interactions are absent. Thus, despite the fact that the single cell dynamics are highly non-linear, the behavior of the cells is pretty much correlated because all cells are invariably undergoing just vegetative growth. In turn, they behave on average like a hypothetical representative cell, which can be adequately
Figure 6.5: Panels (a, b): Transient dynamics of the population mean for the switching from the low ([I_{low}] = 0 µM) to the high state ([I_{high}] = 60 µM) (panel a) and conversely (panel b). Nominal parameter set. For the blue curve (single cell 1), the average values for [O]{\textsubscript{T}} and A/V were taken to be as shown in Figure 6.4a, b, and for the red curve (single cell 2), they were calculated from equations (6.30, 31). Panel (c): As in panel (a) but with $k_{s0MY} = 0.25 \cdot 10^{-1}$ min$^{-1}$, $k_{s1MY} = 0.005 \cdot 10^{-1}$ min$^{-1}$, $\lambda_{MY} = 0.001155$. Panel (d): The distribution of the division times (for comparison with the switching times). Parameters as in Table 5.2 and Table 6.1 unless otherwise noted.

described by a single cell model using good estimates for the average cellular characteristics.

The above results pertain to the time invariant behaviors of the cell population. It is interesting, however, to also compare the dynamical behavior of the single cell and the cell population models.

Figure 6.5a shows such a comparison. The red and blue dashed curves correspond to the predictions of the single cell model and the black curves shows the dynamics of the cell population average. For the blue curve, the average values for the cellular
characteristics were taken to be as shown in panels (a) and (b) of Figure 6.4, and for the red curve, these values were taken to be equal to the estimates of equations (6.30, 31). For panel (a), the single cells models are simulated for \([I_{ex}] = 60 \, \mu M\), using as IC the steady state calculated for \([I_{ex}] = 0 \, \mu M\). For the population model, initially a cell lineage is simulated for \([I_{ex}] = 0 \, \mu M\). At time \(t = 1000\) min, the resulting cell population (\(N_{\text{max}} = 10^4\)) has equilibrated to the stationary distribution that corresponds to \([I_{ex}] = 0 \, \mu M\).

Then, this population is used as an IC for a simulation with \([I_{ex}] = 60 \, \mu M\) in order to observe the dynamics of switching from the low to the high state at the population level. This computational procedure mimics the experimental task of inoculating cells to different environment conditions. Similarly, the dynamics of the opposite switching are simulated (first the cells equilibrate at 60 \(\mu M\) IPTG and then the population is “inoculated” at 0 \(\mu M\) IPTG).

The agreement between the transient behavior of single cell model and the population model is excellent even in the case where the intracellular dynamics are significantly slower than the proliferation rates of the cell as shown in Figure 6.5c, d. Panel (c) shows that for slow \(lacY\) mRNA dynamics the switching from the low to the high state takes approximately 300 min, which is much longer than the 30 min average doubling time. Even so, the single cell model can adequately predict the population average. Therefore, not only the steady states of the single cell agree to a great extend with the stationary cell population averages, the transients exhibit remarkable similarity as well. This similarity can be attributed to the lack of strong inter-cellular coupling or differential cell attributes at the population level.

As a concluding remark to this section, we should note that the comparisons of the
single cell model and the cell population model are purely heuristic. Contrary to the
comparison of the stochastic and deterministic single cell lac operon models (Chapter 5),
where the two models coincide in the limit of large system size, in the present case there
does not exist such a limiting case where the population model reduces to the single cell
model. The latter merely captures a snapshot in the life of some hypothetical
representative cell, but does not take into account the specifics of growth, DNA
duplication and partitioning.

Although Mantzaris (2005b) asserts that the equation for the average of the CPB under
the assumption of no heterogeneity “has the general form of chemically structured
continuum models (Fredrickson 1976)”, his methodological framework has two
problems. First, Mantzaris (2005b) writes the population balance for concentrations,
which are intensive quantities, whereas the population balance only admits contents,
which are extensive. Based on this premise, he assumes that there exists a case where the
NDF can be a Dirac function at all times (no heterogeneity) and then writes the
“continuum” model for this case. However, one expects that for the NDF to be Dirac at
all times, the cells must divide equally and in a synchronized manner. But then, the
average of the NDF will change discontinuously at every division time, and thus its 1st
derivative will be undefined. Hence, an ODE for the average (the alleged continuum
model) cannot even be formulated.

Second, the proposition that the ODE for the average is a structured continuum model
is not valid, since the term that Mantzaris (2005b) perceives as dilution due to cell
proliferation, is in fact irrelevant to the dilution term found in (Fredrickson 1976). The
latter expresses dilution of species concentrations due to cell growth (and thus cell
volume expansion). The former (the negative term in the right hand side of equation 4 in Mantzaris 2005b) expresses the decrease in the average population content due to the binary fission which generates cells with smaller contents than their mother cell.

In summary, the marked deviations between single cell and cell population average behavior shown by Manztaris (2004; 2005a; 2006; 2007) may be artificial, resulting from particular modeling choices or unfair comparison. As we have shown, a detailed modeling of the underlying biological processes yields excellent agreement between single cell and cell population average behavior, provided that reasonably good estimates for the average cell characteristics are used.

6.2 Stochastic Reaction Occurrence

Having analyzed the case where reaction dynamics follow deterministic laws, we will now investigate the case of stochastic reaction occurrence. As discussed in section 5.4 (page 112), the small copy numbers of molecules encountered in this system are expected to result in significant intrinsic stochasticity.

6.2.1 Effect of Stochasticity on Cell Population Behavior

Stochasticity in reaction occurrence results in transcriptional and translational bursts, as shown in panel (a) of Figure 6.6, which depicts the timecourses of the cells in a lineage. Furthermore, panel (b) shows the temporal evolution of the cell population average.

Due to the frequent simulation of the discrete reaction events, the computational expense is much higher in comparison to the deterministic case. Thus, for intermediate to high extracellular IPTG concentrations we can only simulate a limited number of cells
In order to circumvent this problem, a multi-run scheme was employed: for a specific parameter set, a batch of several population simulations was initiated with the same IC but different seeds for the random number generator. The constant number MC technique was used when the population reached a number of cells $N_{\text{cells max}} = 500$. Then, each population of the batch was sampled and finally, all the samples obtained were used to calculate the number density of the batch. Thus, if the batch consisted of $N_{\text{batch}}$ population simulations, and at time $t$ population $i$ consisted of $N_{\text{popul, } i}$ cells, then the final number density would be:

$$n(y, t) dy = \frac{1}{\sum_{i=1}^{N_{\text{batch}}} \sum_{j=1}^{N_{\text{popul, } i}} 1_{\{y_{i,j} \in [y, y + dy]\}}} \cdot \sum_{i=1}^{N_{\text{batch}}} \sum_{j=1}^{N_{\text{popul, } i}} 1_{\{y_{i,j} \in [y, y + dy]\}}$$  \hspace{1cm} (6.32)

where $y$ is the observable of interest (e.g. the total LacY concentration), $y_{i,j}$ is the value of the observable for cell $j$ in population $i$, and $1_{\{E\}}$ is an indicator taking value 1 if event $E$ is true, 0 otherwise. Figure 6.7 shows the results of a simulation batch for the parameter set of Figure 6.6.

Now that we have resolved the issue of simulating a large number of cells, we can
Figure 6.7: Panel (a): the population average calculated from the simulation batch. Panel (b): Number of cells in each of the 20 batches versus time. \([I_{ex}] = 20 \mu M\) and nominal parameter set (Table 5.2, Table 6.1).

elucidate the effect of stochasticity on the cell population heterogeneity by comparing simulations obtained with deterministic reaction dynamics (equation 6.1) versus stochastic reaction occurrence (equation 3.15).

Figure 6.8 shows such comparisons for different extracellular IPTG concentrations.

For low \([I_{ex}]\) stochasticity creates a heavy tail in the NDF, whereas the deterministic simulation exhibits a narrow peak in the total LacY concentration (panel a). This heavy tail can be attributed to the autocatalytic mechanism present in the lac operon system. For intermediate \([I_{ex}]\) values, the deterministic number density is clearly bimodal, in contrast to the stochastic one, which exhibits a much less prominent higher mode. Thus, stochasticity in this case suppresses bistability; yet, a heavy tail indicative of the positive feedback dynamics is still observed. Finally, in panels (c) and (d) the IPTG concentration is high and the deterministic number density is unimodal. However, the stochastic number density appears bimodal, with a sharp peak at total LacY concentration equal to zero and a wide peak at high \([Y]_T\). Stochasticity in this case appears to extend the region where the number density is bimodal and widen the upper mode of the distribution.
Figure 6.8: Comparison of the NDFs for deterministic ("Deter. Rxn") versus stochastic ("Stoch. Rxn") reaction dynamics. For all deterministic simulations the cell population was initiated with 20 cells and $N_{cell_{max}} = 10000$. For all stochastic simulations batches of 20 simulations were run. In all cases sampling was performed at $t = 300$ min. Panel (a): $[I_{ex}] = 10 \, \mu M$. Panel (b): $[I_{ex}] = 20 \, \mu M$. Panel (c): $[I_{ex}] = 30 \, \mu M$. Panel (d): $[I_{ex}] = 50 \, \mu M$. Nominal parameter set (Table 5.2, Table 6.1).

6.2.2 Comparison of Single Cell Probability- and Cell Population Number-Density

Given the extreme computational expense of simulating the population model with stochastic reaction occurrence, we pose the question of whether there is a simpler method for obtaining good approximations for the distributions of phenotypic characteristics. In contrast to the case of deterministic reaction dynamics, here we cannot use an average continuum model. The main reason is that in chemical systems far from the thermodynamic limit, the size influences noise strength. Thus, in the case of stochastic reaction occurrence, growth results in dilution of molar contents and also suppression of
Figure 6.9: Panel (a): Comparison of the single cell probability distribution function with the cell population NDF. For the population distribution, a batch of 10 simulations was run with $N_{\text{cell, max}} = 500$, $[I_{\alpha}] = 30 \mu M$ and nominal parameter set (Table 5.2, Table 6.1), and sampled at $t = 270$ min. For the single cell simulation a cell line was tracked ($N_{\text{cell, max}} = 1$) for $10^5$ min and samples were taken periodically in time with $\Delta t = 10$ min. Panel (b): as in panel (a) with $[I_{\alpha}] = 30 \mu M$ and 100-fold faster $\text{lac}Y$ transcription ($k_{\text{MY} \rightarrow Y} = 50 \text{ min}^{-1}$, $k_{\text{IMY} \rightarrow Y} = 1 \text{ min}^{-1}$) and slower translation ($k_{Y} = 0.3 \text{ min}^{-1}$) and $[I_{\alpha}] = 20 \mu M$. The simulation batch consisted of 20 simulations and was sampled at $t = 250$ min. The cell line was tracked for $10^5$ min of simulated time.

stochastic fluctuations. Consequently, we cannot create a single cell model that accounts for growth with just the incorporation of a dilution term as was done in the deterministic case.

However, we can take a different approach. We have discussed in section 6.1.3 that the cell population dynamics emerge from single cell behavior, and there is a lack of strong intra-cellular coupling mechanisms in our system. Therefore, instead of tracking the NDF in a cell lineage, we could compute the PDF in a cell line. Thus, instead of focusing on the expected number of cells (of the population) that exist in state $z$, we turn our attention to the probability of finding a single cell (of the cell line) at state $z$. Note, however, that we will be comparing the PDF and NDF of intensive quantities, in particular the total LacY concentration.

A comparison of the cell population number density with the single cell probability density for $[Y]_T$ shows a remarkable agreement between the two (Figure 6.9). Such an agreement is obtained for different parameter sets which generate heavy tails or bimodal
distributions. Note that this comparison pertains to stationary conditions and essentially shows that under the assumptions used to build the models, the single cell behavior is indicative of the population behavior. Thus, we can successfully predict the stationary NDF just by simulating a cell line, which is computationally less expensive than simulating the cell lineage.

6.3 Conclusion

In this chapter we incorporated the deterministic and stochastic single cell lac operon models into the cell population framework. We subsequently performed simulations to quantify the phenotypic heterogeneity and made heuristic comparisons between the predictions of the single cell and the cell population models.

These comparisons showed that in the deterministic case, the single cell model gives satisfactory results for the average LacY concentration of the population, even in the case where the single cell dynamics are much slower than the proliferation dynamics. This agreement between the two models was attributed to the lack of strong inter-cellular coupling mechanisms and absence of differential growth rates or unequal partitioning for the cells undergoing vegetative growth.

Furthermore, in the case of the stochastic reaction occurrence, we showed that instead of simulating cell population dynamics we can simulate the dynamics of a cell line. The PDFs thus obtained for the LacY concentration are in excellent agreement to the cell population NDF.
Chapter 7:

Proposed Future Work

In this chapter, we propose future research directions and discuss the possible challenges and impact of the proposed work. This discussion is organized in two parts: the first refers to further development of the framework from an algorithmic perspective, and the second pertains to applications in biological problems.

7.1 Algorithmic Development

7.1.1 Acceleration of the Computational Modules Handling Reaction

The reaction dynamics in the deterministic formulation are simulated using an ODE formulation. In the stochastic formulation, they are simulated as Markov processes. In both cases though, the detailed simulation of reaction events can become extremely computationally expensive, especially since the biochemical reaction networks are large and complex. Overcoming this problem without losing the relevant biological information necessitates the application of rigorous approximations that reduce the computational overhead and admit a stringent control of the resulting error.

In the case of deterministic reaction dynamics, singular perturbation approximations
are excellent candidates for this task. Yet, they frequently suffer from closure problems and intolerable complexity, should one try to perform pencil-and-paper model reduction on a large model. A computational method that can prove useful in this context is the Computational Singular Perturbation (Lam 1993). The development of a module that handles the reaction dynamics using this computational method will not only accelerate the cell population simulations, but also give insight on which reactions are dominant at a given time interval.

In the case of stochastic reaction dynamics, tau-leap schemes, singular perturbation, and quasi steady state or diffusion approximation techniques can be used, depending on the particular magnitudes of the kinetic parameters (Gillespie 2000; Gillespie 2001; Kepler and Elston 2001; Haseltine and Rawlings 2002; Gillespie 2003; Rao and Arkin 2003; Haseltine and Rawlings 2005; Salis and Kaznessis 2005a; Salis and Kaznessis 2005b; Barzel et al. 2007; Samant et al. 2007). In the limiting case, where a species copy number reaches very high numbers, one could also use the deterministic formulation for that species and either a diffusion approximation or the full stochastic simulation of discrete reaction events for the other species. These approximations are expected to result in a marked improvement of computational efficiency. However, as a disadvantage of these approximations, we may note that rarely can one obtain a rigorous error estimate, and therefore validation with the full stochastic simulation is recommended.

7.1.2 Incorporation of Spatial Dimensions and Migration Dynamics

Currently the framework does not take into account space, assuming that all cells exist in a well stirred batch reactor. This assumption would not be valid if one wishes to study migration dynamics of cells on a plate or scaffold. In order to study these dynamics, one
needs to expand the framework first by augmenting the state vector for the single cell, taking into account the position \((x, y)\) of the cell, and then by postulating a law that governs the cell movement. Subsequently, one can develop the corresponding computational module and incorporate it into the framework. In the following, we will discuss the issues related to cell movement.

Postulating the law that governs cell movement depends on the actual biological problem under consideration. A piece-wise linear random walk could be constructed by assuming that the cell moves on a line with constant speed and changes direction after a random time has passed. We will refer to this change of direction as a “turning event”. The interarrival times between the turning events can be exponentially distributed random variables, and the new directions can be uniformly distributed between 0 and \(2\pi\), or they can be dependent on the concentration gradient of a nutrient. Thus, one can allow the cell to bias its random walk towards nutrient-rich areas. Acceleration can also be taken into account in a more detailed description of the cell movement.

An issue that needs to be resolved is the detection of potential collisions between cells with the formulation of rules for either the avoidance of collisions or for the aftermath of a collision. Since the movement of each cell is a linear trajectory between the turning events, one can predict the time of a collision event as the time where the distance
between the positions \((x_i, y_i)\) and \((x_j, y_j)\) of two cells becomes less than a threshold. Then, the cells \(i\) and \(j\) will have to turn: their new directions will be random numbers with a joint density that favors the event of one cell moving away from the other.

7.1.3 Incorporation of Cell Death

The incorporation of cell death is trivial in the framework, since subroutines that remove elements from the data structures storing the event occurrence times have been designed and implemented. What remains is to postulate propensity functions for the death event and incorporate the module that handles the death event. From an algorithmic perspective, the latter procedure is a removal of a cell from the population and a re-labeling of the cells, so that the last cell is being assigned the serial number of the deceased cell.

7.1.4 Incorporation of Cell Cycle for Eukaryotes

The framework has been built for bacterial cells and therefore does not distinguish between cell cycle phases, other than the pre- and post-DNA duplication periods. However, it is readily extensible to the cell cycle phases observed in the eukaryotic cell cycle, namely \(G_1\)-, \(S\)-, \(G_2\) and M-phase.

To this end, the state vector of the cell will have to be augmented by a virtual integer variable ranging from 1 to 4 and expressing the state of the cell. Then, propensity functions for the transition between the cell cycle phases can be postulated. Should one wish to incorporate biological detail in the model, one can build a detailed cell cycle model and formulate those propensities in terms of the concentrations (or contents) of the key regulatory species. This modeling approach will make possible the investigation of
the dependence of division times on the concentration of extracellular nutrients or signaling molecules. Subsequently, distributions for the age of the cells in the population can be calculated. Note that in this proposed approach, the age is an output and not a state variable as in the age-structured population balance approach (Sherer et al. 2008).

7.2 Applications to Biological Problems

7.2.1 Effect of Plasmid Copy Number on Phenotypic Heterogeneity

Most of the current studies about cell population heterogeneity focus on the effect of genetic network structure on the phenotypic variability at the population level. However, in the case that the genetic network is encoded in a plasmid, there are strong indications that the plasmid copy number can significantly affect the phenotypic heterogeneity. Several mechanisms may contribute to this heterogeneity: the extra metabolic burden to the host and the resulting increase in doubling time (Carrier et al. 2000), translational limitations due to insufficient ribosome numbers (Laffend and Shuler 1994b), and finally the variability of plasmid copy numbers and potential loss of plasmid. In applications where a pharmaceutical or biotechnological product is harvested from a cell culture, all these mechanisms are expected to significantly affect the productivity. A model that gives insight of the link between plasmid copy number and the resulting phenotypic distribution (heterogeneity) can thus be used to understand and optimize the process of interest.

In order to build such a model, a network of reactions has to be built that incorporates the dynamics of plasmid synthesis and copy number control, as well as the detailed processes of protein production through transcription and translation. The subpart of the
reaction network that handles plasmid copy number regulation can be built using the specific biological information pertinent to the origin of replication for the plasmid of interest (Khan 1997; Actis et al. 1999; Solar and Espinosa 2002). For the modeling of transcription and translation, the dynamics of the RNA polymerase and ribosomal processing have to be taken into account, by incorporating the known biology for the regulation and function of these species (Dennis and Nomura 1975; Ralling et al. 1985; Nomura 1999; Klumpp and Hwa 2008). Finally, a metabolic “cost” has to be associated with each transcription and translation event; these parameters will influence the growth rate of the cell in order to model the metabolic burden that a high copy plasmid puts on the cell.

After the genetic network has been built, it can be incorporated into the cell population framework and used to simulate the cell population distributions for a protein encoded in high or low copy plasmids. Subsequently, key measures, such as the mean and the variance, can be computed in order to quantify the effect of plasmid copy number on the protein production levels and the degree of heterogeneity. Furthermore, the distributions for the cell proliferation rates (which are proportional to the inverse of the doubling times) can also be obtained. It is expected that the ratio between protein production versus cell proliferation rate will go through a maximum, at which the productivity of the desired product is maximized.

7.2.2 Tissue Growth in Scaffolds

Tissue growth processes and their relation to population dynamics have been studied using phenomenological models (Cheng et al. 2006; Youssef et al. 2007). In these models, cell motility is handled by a cellular automaton formulation and the parameters
controlling cell movement are inferred from experimental data or measured directly. Cell division occurs after a random time, whose probability density depends on the cell type. Simulation of the process of tissue growth is done until confluence is reached, namely when the scaffold is full of cells, and thus further migration or division is impossible.

These studies have given significant insight on tissue growth under different conditions, such as in cases where nutrients are limited or when two cell subpopulation with different motilities are migrating into the scaffold. Yet, they lack the biological detail that links the intracellular mechanisms the observed phenotype. In particular, the probability density for the division times is taken to be fixed, and thus the effect of extracellular environment on proliferation rate is not incorporated into the model.

This problem can be addressed using the framework developed in Chapter 3 with the extensions proposed in sections 7.1.2, 7.1.3 and 7.1.4. Hence, one will first need to build a detailed cell cycle model linked to environmental factors crucial for cell proliferation. These factors may be the availability of nutrients and growth hormones and also the confluence of the tissue within the neighborhood of the cell. Then, one will have to incorporate this model into the framework and formulate a propensity function for the division. Thus, the division times will now be an output of the model, incorporating biological information that links the cell cycle physiology to extracellular conditions. At a next level of detail, cell death can be taken into account by the formulation of a corresponding propensity function resulting in high death rates when nutrient depletion occurs. Furthermore, cell motility will now be handled by a continuous model, which is more realistic then the cellular automaton, since it allows for greater freedom in the movement of cell and can incorporate the dynamics of motion if acceleration is taken into
account.

The resulting model can be used to gain insight into the effect of extracellular environment on the cell proliferation rates and the final degree of confluence in the scaffold. It has been experimentally shown that inside a crowded scaffold, nutrient diffusion is limited, and cell movement is impeded, a phenomenon that results in the death of cells at the interior of the scaffold. In a less crowded scaffold, one expects that the cells existing at the interior will migrate towards the exterior of the scaffold towards more nutrient-rich areas. This process may reduce congestion at the interior and finally result in a balanced proliferation of cells throughout the scaffold. Insights obtained form such simulations may be used, for example, to guide the design of strategies for wound healing in the field of regenerative medicine.

7.3 Conclusion

The modeling framework that was developed in this study is highly customizable and extensible. The attributes of the framework presented and the multitude of potential extensions open exciting directions for future research. Thus, in this last chapter is suggested discussed the development of modules that can be incorporated to the framework in order to study the effect of plasmid copy number on the phenotypic heterogeneity and the consequences of stochasticity on cell migration and death during tissue growth.
Appendices

Appendix I

Mapping a uniform deviate to a random number of known continuous distribution

We need a rule that will allow us to map a uniform deviate $u$ to a random number $x$ with $\Omega = \mathbb{R}_0^+$, that follows any desired continuous PDF $f(x)$. First we define the cumulative density function:

$$F(x) := \int_0^x f(x') \, dx' \quad (AI.1)$$

Then, assuming that $F(x)$ is a one to one mapping, the random number $\chi$ defined as:

$$\chi := F^{-1}(u) \quad (AI.2)$$

follows the probability density $f(x)$. Thus, in order to find the random number $\chi$, one has to essentially solve the following equation:

$$u - F(\chi) = 0 \quad (AI.3)$$

In the cases where this is impossible (i.e. when $F^{-1}$ does not exist) or impractical, there are other rules that can be used based on acceptance-rejection techniques. We will not go into further details since readily available subroutines can be found for the generation of frequently used random numbers that (Numerical Recipes in Fortran, etc).

Application: for finding the random waiting time needed for a reaction event:

$$f(\tau_r) = p_{\text{some ran}}(t + \tau_r, t, X, \Phi(X, V(t), \tau_r)) =$$

$$a_r(X, \Phi(X, V(t), \tau_r)) \cdot \exp \left[ - \int_0^{\tau_r} a_r(X, \Phi(X, V(t), \tau')) \, d\tau' \right] \quad (AI.4)$$
\[ F(\tau) = \int_{0}^{\tau} a_t(\mathbf{X}, \Phi(\mathbf{X}, V(t), \tau)) \cdot \exp \left[ -\int_{0}^{\tau} a_t(\mathbf{X}, \Phi(\mathbf{X}, V(t), \tau')) dt' \right] dt \]

\[ = -\int_{0}^{\tau} \frac{d}{dt} \left[ \exp \left[ -\int_{0}^{\tau} a_t(\mathbf{X}, \Phi(\mathbf{X}, V(t), \tau)) dt' \right] \right] dt \]

\[ = 1 - \exp \left[ -\int_{0}^{\tau} a_t(\mathbf{X}, \Phi(\mathbf{X}, V(t), \tau)) dt' \right] \]  \hspace{1cm} (AI.5)

Thus, we need to solve:

\[ \int_{0}^{\tau} a_t(\mathbf{X}, \Phi(\mathbf{X}, V(t), \tau')) dt' - \ln \left[ (1 - u)^{-1} \right] = 0 \]  \hspace{1cm} (AI.6)

**Mapping a uniform deviate to a random number of known discrete distribution**

We need a rule that will allow us to map a uniform deviate \( u \) to a random number \( \eta \) with \( \Omega = N_0^* \), that follows any desired continuous PDF \( f_\eta \). The random number \( \eta \) defined as:

\[ \eta = \inf \left\{ u \leq \sum_{i=0}^{n'} f_i \right\} \]  \hspace{1cm} (AI.7)

Thus, in order to find the random number \( \eta \), one has to essentially loop through the partial sums of \( f_\eta \) and find the first \( \eta \) such that \( u \) is less than the partial sum \( f_0 + \ldots + f_\eta \).
Appendix II

Normalization condition for the Janossy density

In order to find the proper normalization condition for \( J \), we make use of the abstract idea of a multiset, which is similar to a set with elements, but with the difference that each element can appear multiple times in a multiset. Let \( M \) be a multiset and \( N \) be the set of elements of \( M \). Apparently the cardinality of \( M \) is greater or equal to that of \( N \), since we count every element of \( M \) as many times as it appears, whereas in \( N \) each element is unique. We thus define the following function:

\[
\mathcal{G}: M \mapsto N \\
N = \mathcal{G}(M) = \{\text{set of elements of the multiset } M\}
\]  

(AII.1)

Furthermore, let \( \text{mult}(e_i) \) be the multiplicity of each element of the multiset \( M \).

Example: the multiset \( M = \{a, a, c, b, a, c, b, d\} \) has cardinality 8 and the corresponding set is \( N = \{a, c, b, d\} \) with cardinality 4.

Further, the multiplicity of each element of \( N \) in \( M \) is \( \text{mult}(a) = 3 \), \( \text{mult}(b) = 2 \), \( \text{mult}(c) = 2 \) and \( \text{mult}(d) = 1 \). Apparently, the sum of multiplicities is equal to the cardinality of \( M \).

We now pose the question: “how many orderings (equivalently: permutations) can we create from the elements of a multiset \( M \)?” If we were talking about a set \( N \), the answer is \( \text{card}(N)! \), the factorial of the cardinality of \( N \). However, in the case of a multiset, the number of possible orderings will be reduced due to fact that if we permute the same elements we end up with the exact same ordering. Thus, the number of possible orderings will be:

\[
\text{ords}(M) = \frac{\text{card}(M)!}{\prod_{i=1}^{\text{card}(\mathcal{G}(M))} \text{mult}(\mathcal{G}(M)_i)}
\]  

(AII.2)
For example, for the set introduced above $M = \{a, a, c, b, a, c, b, d\}$, we can produce 1680 different orderings:

$$\text{ords}(a, a, c, b, a, c, b, d) = \frac{8!}{3! \cdot 2! \cdot 2! \cdot 1!} = 1680$$

Now, let us go back to the cells in a population. We distinguish between the following cases: (i) the states of the cells are vectors of discrete variables, (ii) the states of the cells are vectors of continuous variables, (iii) the states of the cells are vectors of continuous and discrete variables.

**Case (i): discrete**

In order to derive the normalization condition for the Janossy density we will have to sum over all possible number of cells in the cell population and over all the discrete space of possible cell states. We observe though that in this last sum over states, some terms are repetitive: $J_3(z_1,z_1,z_2)$ is the same as $J_3(z_1,z_2,z_1)$ but it will be summed twice. Thus, we will have to use weighting factors equal to the inverse of the number of possible orderings of $(z_1,z_1,z_2)$. In general the normalization condition will be:

$$\sum_{v \geq 0} \left\{ \sum_{z_1} \ldots \sum_{z_v} \frac{1}{\text{ords}(z_1, \ldots, z_v)} \cdot J_v(z_1, \ldots, z_v; t) \right\} = 1 \quad \text{(AII.3)}$$

**Case (ii): continuous**

For the continuous case we will have to sum over all possible number of cells in the population and integrate over all possible states. In this case repetitions will still occur, for example $J_3(z_1,z_1,z_2)$ and $J_3(z_1,z_2,z_1)$ are the same. However, the manifold $\{x_1 = z_1, x_2 = z_1, x_3 = z_2\}$ is a 2-dimensional surface in a 3-dimensional space, and thus has zero Lebesgue measure. In general, integration of the Janossy density over the lower dimensional manifolds that occur due to repetitions, does not contribute to the overall
integral, unless the density contains terms with delta functions over such manifolds.

Thus, we can write:

$$\sum_{\nu \geq 0} \left\{ \int \cdots \int \frac{1}{\nu!} J_{\nu}(z_1, \ldots, z_\nu; t) \, dz_1 \cdots dz_\nu \right\} = 1 \quad \text{(AII.4)}$$

This equation provides the normalization condition used in Ramkrishna (2000). Of course one realizes that, if the density contains Dirac delta terms, which extract a lower dimensional manifold, then the above normalization condition is no longer true for instance consider:

$$J_3(z_1, z_2, z_3; t) = \delta(z_1 - z_2) + \delta(z_1 - z_3) + \delta(z_2 - z_3) - 2 \cdot \delta(z_1 - z_2) \cdot \delta(z_2 - z_3) \quad \text{(AII.5)}$$

This expression is a valid Janossy density since it is invariant under permutations; notice also that it is “sufficiently continuous”: if we are moving in the plane $z_1 = z_2$ then for $z_2 \neq z_3$ we get one Dirac delta term, and for $z_1 = z_2 = z_3$ we also get one delta term as desired.

In this case the correct normalization condition is (suppose $J_{\nu} = 0 \forall \nu \neq 3$):

$$\int \cdots \int \frac{1}{3!} J_{3}(z_1, \ldots, z_3; t) \, dz_1 \cdots dz_3 = \int \cdots \int \frac{2!}{3!} J_{3}(z_1, \ldots, z_3; t) \, dz_1 \cdots dz_3 = 1 \quad \text{(AII.6)}$$

Of course this case is exceptional and will not be of concern from now on.

**Case (iii): discrete-continuous**

In our case the state vector for each cell contains both discrete and continuous variables. The argument of the continuous case can also be used here, since, for two cells to be at the exact same state, all of their state variables, including the continuous one, will have to be equal. But the requirement for the continuous variable to have the same value for both cells essentially defines a lower dimensional manifold, integration over which does not contribute to the overall integral. Thus, the normalization condition will be:
\[ \sum_{v} \left\{ \sum_{x_1} \sum_{x_v} \left[ \frac{1}{v!} \cdot J_v((x_1, v_1), ..., (x_v, v_v); t) \right] dV_v \ldots dV_1 \right\} = 1 \] (AII.7)
Appendix III

Alternative Reaction Paths and the Origin of Equation (5.11)

Equation (5.11) essentially originates from Hess’s law for the Gibbs free energy applied to the set of chemical equilibria (5.5 - 10). Hess’s law expresses the conservation of energy, denoted as state function $\Delta G$, regardless of the path through which it is to be determined. In our case, there exist three linearly dependent chemical equilibria:

\[
\begin{align*}
R_2 + O & \rightleftharpoons _{K_1}^{K_1} R_2O & (5.5, 6) \\
2I + R_2 & \rightleftharpoons _{K_2}^{K_2} I_2R_2 & (5.7, 8) \\
2I + R_2O & \rightleftharpoons _{K_3}^{K_3} I_2R_2 + O & (5.9, 10)
\end{align*}
\]

where $K_i = \frac{k_i}{k_{-i}}$ denotes the equilibrium constant of reversible reaction $i$. Linear dependence in the equilibria means that there are two "paths" for going from one state to another:

path 1: \[ R_2 + O + 2I \xrightarrow{\text{fwd 1}} R_2O + 2I \xrightarrow{\text{fwd 3}} I_2R_2 + O \] (AIII.1)

path 2: \[ R_2 + O + 2I \xrightarrow{\text{fwd 2}} R_2I_2 + O \] (AIII.2)

where fwd (bwd) $i$ means that reversible reaction $i$ is taking place in the forward (backward) direction. Since the initial and final state for both paths are the same, it follows the Gibbs free energy for the three equilibria satisfies the equation:

\[
\begin{align*}
\Delta G_1 + \Delta G_3 &= \Delta G_2 \\
R \cdot T \cdot \ln (K_1) + R \cdot T \cdot \ln (K_3) &= R \cdot T \cdot \ln (K_2) & \Rightarrow \\
K_1 \cdot K_3 &= K_2
\end{align*}
\]

The latter equation is essentially equation (5.11).
Appendix IV

Limiting Relations for the Deterministic Model for High Inducer Concentrations

After setting the derivatives equal to zero, equations (5.28 - 36) can be simplified to the following system. From equation (5.28):

\[
[M_R] = \frac{k_{aMR}}{\lambda_{MR}}
\]  
(AIV.1)

From equation (5.29):

\[
[R_2] = -\frac{k_{2R}}{2 \cdot k_{-2R}} \cdot [M_R] + \frac{k_{2R}}{k_{-2R}} \cdot [R] + \frac{\lambda_R}{2 \cdot k_{-2R}} \cdot [R]
\]  
(AIV.2)

From the linear combination of equations (5.30) – (5.31) + (5.33):

\[
[I_2R_2] = \frac{k_{2R}}{\lambda_{12R2}} \cdot [R]^2 - \frac{k_{-2R} + \lambda_{R2}}{\lambda_{12R2}} \cdot [R_2]
\]  
(AIV.3)

From equations (5.31) – (5.33):

\[
0 = -k_r \cdot [R_2] \cdot [O] + k_{-r} \cdot ([O]_T - [O]) - k_{dr} \cdot [I_2] \cdot [I]^2 + (k_{-dr} + \lambda_{12R2}) \cdot [I_2R_2]
\]  
(AIV.4)

From equation (5.31):

\[
0 = -k_r \cdot [R_2] \cdot [O] + k_{-r} \cdot ([O]_T - [O]) + k_{dr} \cdot ([O]_T - [O]) \cdot [I]^2 + ...
\]  
(AIV.5)

From equation (5.34):

\[
[M_Y] = \frac{k_{s0MY}}{\lambda_{MY}} \cdot [O] + \frac{k_{aMY} - k_{s0MY}}{\lambda_{MY}} \cdot [O]
\]  
(AIV.6)

From equation (5.35):

\[
[M_Y] = \frac{\lambda_{Y_{lex}}}{k_{Y_{lex}}} \cdot [Y_{lex}] + \frac{\lambda_Y}{k_{Y}} \cdot [Y]
\]  
(AIV.7)

From the linear combination of equations (5.32) + (5.33):
\[
[Y_L_{\text{en}}] = \frac{k_1}{k_n + \lambda_{\text{en}}} \cdot ([I] - [I_{\text{ex}}])
\]  
\[\text{(AIV.8)}\]

From equation (5.36) using (AIV.8):

\[
[Y] = \left(\frac{k_n + k_p + \lambda_{\text{en}}}{k_p}\right) \cdot \left(\frac{[I]}{[I_{\text{ex}}]} - 1\right)
\]  
\[\text{(AIV.9)}\]

We wish to investigate the steady state limiting behavior of the system at high induction levels thus take:

\[
[I_{\text{ex}}] = \frac{\sigma}{\varepsilon}
\]  
\[\text{(AIV.10)}\]

and we introduce asymptotic expansions with respect to \(\varepsilon\) for all the species. Since \([M_R]\) is constant:

\[
[M_R] = [M_R]_0
\]  
\[\text{(AIV.11)}\]

and since \([O]\) is bounded by \([O]\)_{T}:

\[
[O] = [O]_0 + [O]_1 \cdot \varepsilon + [O]_2 \cdot \varepsilon^2 + ...
\]  
\[\text{(AIV.12)}\]

For all the other species we do not know whether they do or do not have a growing component of \(O(\varepsilon^{-1})\), and thus we take the following expansions:

\[
[R] = \frac{[R]_0}{\varepsilon} + [R]_1 + [R]_1 \cdot \varepsilon + [R]_2 \cdot \varepsilon^2 + ...
\]  
\[\text{(AIV.13)}\]

\[
[R_2] = \frac{[R_2]_0}{\varepsilon} + [R_2]_1 + [R_2]_2 \cdot \varepsilon + [R_2]_3 \cdot \varepsilon^2 + ...
\]  
\[\text{(AIV.14)}\]

\[
[I_2R_2] = \frac{[I_2R_2]_0}{\varepsilon} + [I_2R_2]_1 + [I_2R_2]_2 \cdot \varepsilon + [I_2R_2]_3 \cdot \varepsilon^2 + ...
\]  
\[\text{(AIV.15)}\]

\[
[I] = \frac{[I]_0}{\varepsilon} + [I]_1 + [I]_2 \cdot \varepsilon + [I]_3 \cdot \varepsilon^2 + ...
\]  
\[\text{(AIV.16)}\]
\[ [M_Y] = \frac{[M_Y](-1)}{\varepsilon} + [M_Y]_0 + [M_Y](0) \cdot \varepsilon + [M_Y](2) \cdot \varepsilon^2 + ... \quad \text{(AIV.17)} \]

\[ [Y] = \frac{[Y](-1)}{\varepsilon} + [Y]_0 + [Y](0) \cdot \varepsilon + [Y](2) \cdot \varepsilon^2 + ... \quad \text{(AIV.18)} \]

\[ [Y_{Ie}] = \frac{[Y_{Ie}](-1)}{\varepsilon} + [Y_{Ie}]_0 + [Y_{Ie}(0)](0) \cdot \varepsilon + [Y_{Ie}(2)] \cdot \varepsilon^2 + ... \quad \text{(AIV.19)} \]

Now, if we substitute the asymptotic expansions to the steady state equations (AIV.1 - 8) we find the following:

\[ \mathcal{O}(\varepsilon^{-1}) : \]

From equations (AIV.2) and (AIV.3):

\[ [R_2](-1) = \frac{k_{2R}}{k_{-2R}} \cdot 2 \cdot [R](-1) \cdot [R]_0 + \frac{\lambda_{R}}{2 \cdot k_{-2R}} \cdot [R](-1) \quad \text{(AIV.20)} \]

\[ [I_2R_2](-1) = \frac{k_{2R}}{\lambda_{12R2}} \cdot 2 \cdot [R](-1) \cdot [R]_0 - \frac{k_{2R} + \lambda_{R2}}{\lambda_{12R2}} \cdot [R_2](-1) \quad \text{(AIV.21)} \]

From (AIV.20) and (AIV.21) it follows that:

\[ 0 = \frac{\lambda_{R}}{2} \cdot [R](-1) + \lambda_{R2} \cdot [R_2](-1) + \lambda_{12R2} \cdot [I_2R_2](-1) \quad \text{(AIV.22)} \]

But \([R](-1), [R_2](-1)\) and \([I_2R_2](-1)\) must all be non-negative otherwise the \([R], [R_2]\) and \([I_2R_2]\) will diverge to \(-\infty\). Thus, necessarily:

\[ [R](-1) = 0 \quad \text{(AIV.23)} \]

\[ [R_2](-1) = 0 \quad \text{(AIV.24)} \]

\[ [I_2R_2](-1) = 0 \quad \text{(AIV.25)} \]

Furthermore from (AIV.4) and (AIV.5) using (AIV.23 - 25):
\[ 0 = \left[R_2\right]_{(0)} \cdot \left[I\right]_{(-1)} \cdot \left[I\right]_{(0)} \]  (AIV.26)

\[ 0 = \left[(O)_{(0)} - [O]_{(0)}\right] \cdot [I]_{(-1)} \cdot [I]_{(0)} \]  (AIV.27)

And from equations (AIV.6, 7) since [O] and [O]_{(0)} are bounded:

\[ [M_Y]_{(-1)} = 0 \]  (AIV.28)

\[ [M_Y]_{(-1)} = \frac{\lambda_{Y_{ex}}}{k_x} \cdot [Y_{I_ex}]_{(-1)} + \frac{\lambda_Y}{k_y} \cdot [Y]_{(-1)} \]  (AIV.29)

From (AIV.29) since [Y_{I_ex}]_{(-1)} and [Y]_{(-1)} must be non-negative it immediately follows that:

\[ [Y]_{(-1)} = 0 \]  (AIV.30)

\[ [Y_{I_ex}]_{(-1)} = 0 \]  (AIV.31)

But from (AIV.8):

\[ [Y_{I_ex}]_{(-1)} = \frac{k_x}{k_x + \lambda_{Y_{ex}}} \cdot \left([I]_{(-1)} - \sigma\right) \]  (AIV.32)

which means that:

\[ [I]_{(-1)} = \sigma \]  (AIV.33)

Therefore, we deduced that none of the concentrations with the exception of [I] has any growing components, in other words they remain bounded as [I_{ex}] tends to infinity.

On the contrary [I] can always enter the cell by means of free diffusion through the membrane, and thus it grows in the same order of magnitude as [I_{ex}].

\[ O(1) \]

Trivially from (AIV.1):
\[
[M_R]_{(0)} = \frac{k_{\text{MR}}}{\lambda_{\text{MR}}}
\] (AIV.34)

From equation (AIV.9) using (AIV.33):

\[
[Y]_{(0)} = 0
\] (AIV.35)

From equation (AIV.8):

\[
[Y_{\text{ex}}]_{(0)} = \frac{k_{\text{I}}}{k_{\text{f}} + \lambda_{Y_{\text{ex}}}} \cdot [I]_{(0)}
\] (AIV.36)

and therefore from (AIV.7):

\[
[M_Y]_{(0)} = \frac{\lambda_{\text{Yex}}}{k_{\text{fY}}} \cdot [Y_{\text{ex}}]_{(0)}
\] (AIV.37)

Also from (AIV.6):

\[
[M_Y]_{(0)} = \frac{k_{s0\text{MY}}}{\lambda_{\text{MY}}} \cdot [O]_{r} + \frac{k_{s1\text{MY}} - k_{s0\text{MY}}}{\lambda_{\text{MY}}} \cdot [O]_{(0)}
\] (AIV.38)

Now in equation (AIV.38) \(k_{s1\text{MY}} - k_{s0\text{MY}} > 0\) since the rate at which the free promoter transcribes the LacY gene is faster than that of the repressed promoter; thus \([M_Y]_{(0)}\) is positive. From the previous equations (AIV.36, 37) it follows that \([Y_{\text{ex}}]_{(0)}\) and \([I]_{(0)}\) are also positive. Thus, from equations (AIV.26) and (AIV.27) we deduce that:

\[
[R_2]_{(0)} = 0
\] (AIV.39)

\[
[O]_{(0)} = [O]_{r}
\] (AIV.40)

The total LacY at high induction levels will be equal to \([Y_{\text{ex}}]_{(0)}\) which can be calculated to be:

\[
[Y_{\text{ex}}]_{(0)} = \frac{k_{\text{fY}} \cdot k_{s1\text{MY}}}{\lambda_{Y_{\text{ex}}} \cdot \lambda_{\text{MY}}} \cdot [O]_{r}
\] (AIV.41)

The total repressor concentration at high induction levels can also be calculated. From
equations (AIV.1, 2) and (AIV.3):

$$0 = 2 \cdot k_{2R} \cdot [R]_{(0)}^2 + \lambda_R \cdot [R]_{(0)} - k_{sr} \cdot \frac{k_{AMR}}{\lambda_{MR}}$$ \hspace{1cm} (AIV.42)

$$[l_2R_2]_{(0)} = \frac{k_{2R}}{\lambda_{12R2}} \cdot [R]_{(0)}^2$$ \hspace{1cm} (AIV.43)

Consequently, the discriminant and the solution for $[R]_{(0)}$ are given as:

$$\Delta = \lambda_R^2 + 8 \cdot k_{2R} \cdot k_{sr} \cdot \frac{k_{AMR}}{\lambda_{MR}} > \lambda_R^2$$ \hspace{1cm} (AIV.44)

$$[R]_{(0)} = \frac{-\lambda_R + \sqrt{\Delta}}{4 \cdot k_{2R}}$$ \hspace{1cm} (AIV.45)

From equations (AIV.39, 40, 43) and (AIV.45) it follows that the total dimeric repressor concentration at high induction is:

$$\lim_{[I_a] \to \infty} [R_2]_T = [R_2]_{(0)} + [R_2O]_{(0)} + [l_2R_2]_{(0)} = \frac{(\sqrt{\Delta} - \lambda_R)^2}{16 \cdot \lambda_{12R2} \cdot k_{2R}}$$ \hspace{1cm} (AIV.46)

Therefore we have reached the following conclusions:

(i) The concentrations of all intracellular species except $[I]$ always remain bounded.

For high induction levels:

(ii) the concentration of intracellular IPTG, $[I]$, grows at the rate as the extracellular, $[I_{ex}]$, due to free diffusion of IPTG through the cell membrane.

(iii) the concentration of the LacY permease, $[Y]$ tends to zero, as more and more of the permease is found in the bound form $YI_{ex}$.

(iv) the concentration of the free repressor $[R_2]$ also tends to zero as more and more of the repressor is found in the bound form $l_2R_2$.

(v) all of the operator content is found in the free (non-repressed) form.
Appendix V

MC Simulations and Calculation of Statistics

Kinetic MC Simulations were performed using the Direct Method of Gillespie's algorithm. Due to the ergodic properties of the process under consideration, sampling over the ensemble is equivalent to sampling over time (van Kampen 1992, p. 190 - 191). Thus, long simulations ($10^4 - 10^6$ min) were performed and samples were taken over constant $\Delta t$. Subsequently, the statistical properties of interest were estimated from these samples.

Thus, for the calculation of the marginal PMF for species $X$, $\Delta t$ was taken $5 \cdot 10^{-2}$ min and the following estimator was used:

$$\hat{p}_X(j) = \frac{1}{N} \cdot \sum_{i=1}^{N} \mathbf{1}_{\{X_i=j\}}$$  \hspace{1cm} (AV.1)

where $p_X(j)$ is the probability of observing $j$ molecules of species $X$ ($j = 0, 1, \ldots$), $N$ is the number of samples and $\mathbf{1}_{\{E\}}$ is an indicator taking value 1 if event $E$ is true, 0 otherwise. The probabilities $p_j$ were subsequently used in the calculation of the mean $\mu_X$ and variance $\sigma_X^2$ of the number of molecules of $X$. If $j_{\text{max}}$ denotes the maximum number of molecules of $X$ within our samples, then:

$$\hat{\mu}_X = \sum_{j=0}^{j_{\text{max}}} p_j \cdot j$$  \hspace{1cm} (AV.2)

$$\hat{\sigma}_X^2 = \sum_{j=1}^{j_{\text{max}}} p_j \cdot (j - \hat{\mu}_X)^2$$  \hspace{1cm} (AV.3)

The autocovariance function $\kappa(\tau)$ can be calculated at time lags $\tau$ that are integer multiples of the $\Delta t$ used in the sampling of the stochastic path which was taken $\Delta t =$
5\cdot10^{-1} \text{ min. The estimator that was used is:}

\hat{k}_X (k \cdot \Delta t) = \frac{1}{N - k - 1} \left( \left\{ \sum_{i=1}^{N} X_i \cdot X_{i-k} - \frac{1}{N-k} \cdot \sum_{i=1}^{N} X_i \cdot \sum_{i=1}^{N} X_{i-k} \right\} \right) \quad \text{(AV.4)}

Note that the above estimators for the mean variance and autocovariance refer to the numbers of molecules of species X. In the comparisons with the deterministic model, we need to work with concentrations because the deterministic results are expressed in concentrations. For example we need to calculate $\hat{\mu}_{[X]}$, the mean concentration $[X]$. Thus, since:

$$[X] = \frac{X}{N_A \cdot V_{E.coli}} \quad \text{(AV.5)}$$

it follows that:

$$\hat{\mu}_{[X]} = \frac{\hat{\mu}_X}{N_A \cdot V_{E.coli}} \quad \text{(AV.6)}$$

$$\hat{\sigma}_{[X]} = \frac{\hat{\sigma}_X}{(N_A \cdot V_{E.coli})^2} \quad \text{(AV.7)}$$

$$\hat{k}_{[X]} (k \cdot \Delta t) = \frac{\hat{k}_X (k \cdot \Delta t)}{(N_A \cdot V_{E.coli})^2} \quad \text{(AV.8)}$$

For the calculation of the First Passage Times, the following methodology was developed: first, two reference states were defined, the high state e.g. for species X having a copy number of 350, and the low state for $X = 5$. During the simulation, or in post-processing mode, the time instances for which the system exists in the low and the high reference state are stored ordered in two separate vectors, $t_{low}$ and $t_{high}$ (these time instances are multiples of a timestep $\Delta t$ which was taken $10^{-2}$ min). Then, in order to find the distribution of the First Passage Time from the low to high reference state, samples
are obtained as follows: let $t_{\text{low}[i]}$ be an element of vector $t_{\text{low}}$. The code locates the first element $t_{\text{high}[j]}$ that satisfies $t_{\text{high}[j]} > t_{\text{low}[i]}$. Thus $(t_{\text{high}[j]} - t_{\text{low}[i]})$ is one sample. Subsequent samples are taken for different $i$ values (by scanning vector $t_{\text{low}}$). With the same methodology the First Passage Time from the high to the low reference state can also be computed. The Mean First Passage Time was estimated as the average of the obtained samples and the variance as the sample variance.
Appendix VI

Derivation of the Propensities for \( I_{\text{ex-related processes}} \)

Consider a cell of volume \( V \) and membrane area \( A \) (\( A \) and \( V \) both being constant). The extracellular IPTG molecules that continuously "hit" the membrane of the cell can either be absorbed into the intracellular space or they can bounce back without leaving the extracellular space. Let us first deal with free IPTG diffusion through the membrane. We approximate the cell membrane with a planar surface at small spatial scales (essentially, we are dealing with a differential area \( \delta A \) of the membrane). We also approximate the IPTG molecule with a sphere. Then, an IPTG molecule will hit the membrane if the distance between the center of the molecule and the surface reduces to half the diameter of the IPTG molecule \( d_{\text{IPTG}/2} \). Thus, the differential area is covered by a "collision volume" (see Figure AVI.1):

\[
\delta V_{\text{cell}} = \delta A \cdot \left(-u_{\text{IPTG}} \cdot \hat{n}_A \right) \cdot 1_{\left[-\delta A, \delta A\right]} \cdot \delta t
\]  

(AVI.1)

where \( 1_{\left[\delta A\right]} \) is unity when \( \delta A \) is true and zero otherwise (the corresponding term in AVI.1 is unity only when the molecule approaches the membrane); \( u_{\text{IPTG}} \) is the velocity of the IPTG molecule relative to the membrane and \( \hat{n}_A \) is the normalized vector perpendicular to the membrane surface with direction towards the outer of the cell. \((\hat{x}, \hat{y})\) denotes the dot product of vectors \( \hat{x} \) and \( \hat{y} \). Thus, if the IPTG molecule is inside the collision
volume then it will hit the membrane at the next $\delta t$ time interval. Since IPTG molecules are uniformly distributed in the extracellular space, it follows that the average probability of one IPTG molecule hitting the membrane is:

$$\left\langle \frac{\delta V_{\text{coll}}}{V_{\text{ex}}} \right\rangle = A \cdot u_m \cdot \frac{\delta t}{V_{\text{ex}}}$$ \hspace{1cm} (AV1.2)

where $A$ is surface area of the whole membrane and $u_m$ is the mean velocity with which IPTG molecules approach the membrane, defined as:

$$u_m = \int \int (-y, n_A) \cdot 1_{\{y_{\text{PTG}}, n_A > 0\}} \cdot p_{y_{\text{PTG}}} (y) dy dA$$ \hspace{1cm} (AV1.3)

If the number of IPTG molecules that hit the membrane per unit time is high, let $f_t$ denote the probability of a successful pass-through given that a hit has occurred.

$$A \cdot u_m \cdot f_t \cdot \frac{\delta t}{V_{\text{ex}}} = \text{average probability, to first order in } \delta t, \text{ that a specific IPTG molecule will successfully pass through the membrane in the next time interval } dt$$ \hspace{1cm} (AV1.4)

Furthermore, there exist $I_{\text{ex}}$ molecules uniformly distributed in the extracellular space:

$$I_{\text{ex}} = V_{\text{ex}} \cdot [I_{\text{ex}}] \cdot N_A$$ \hspace{1cm} (AV1.5)

where $I_{\text{ex}}$ denotes number of molecules, $[I_{\text{ex}}]$ concentrations, $V_{\text{ex}}$ the extracellular volume and $N_A$ Avogadro's number. Thus, if we consider the passing-though of one IPTG molecule as a reaction occurring in the intracellular space that leads to the production of one IPTG molecule therein (assuming the intracellular space being well stirred), then the average reaction rate (in units of molecules produced per time) will be $c'_t \cdot \langle I_{\text{ex}} \rangle \cdot \delta t$ where the stochastic reaction constant (having units of inverse time) is:

$$c'_t = \frac{A \cdot u_m \cdot f_t}{V_{\text{ex}}}$$ \hspace{1cm} (AV1.6)
Note that $c'_t$ is inversely proportional to the extracellular volume. This is expected: for a fixed amount of molecules in the extracellular space, the average number of molecules that pass through the membrane will be higher if the extracellular space is smaller. Now, the deterministic “reaction rate constant equivalent” $k_t$ is conventionally defined to be the average stochastic reaction rate (in mols per time) per unit intracellular volume $V$ divided by the average concentration of the “reactant” (namely the IPTG in the extracellular space):

$$k_t = \frac{1}{V \cdot N_A} \frac{c'_t \cdot \left< I_{ex} \right> \cdot V_{ex} \cdot N_A}{\left< I_{ex} \right>} \Rightarrow k_t \cdot \frac{V}{V_{ex}} = c'_t \quad (AVI.7)$$

For our study, we consider the limiting case where the extracellular volume tends to infinity, and thus $c'_t$ tends to zero. Further, since the IPTG concentration (intensive quantity) remains constant, the number of IPTG molecules in the extracellular volume also tends to infinity, but the product $c'_t \cdot \left< I_{ex} \right> = A \cdot u_m \cdot f_t \cdot \left< I_{ex} \right> \cdot N_A$ remains constant. In the bulk, there is no distinction between $\left< I_{ex} \right>$ and $[I_{ex}]$, and thus we define the following stochastic reaction constant:

$$c_t = c'_t \cdot \left< I_{ex} \right> = k_t \cdot \frac{V}{V_{ex}} \cdot \left< I_{ex} \right> = k_t \cdot V \cdot N_A \cdot [I_{ex}] \quad (AVI.8)$$

which is precisely what was used for the reaction describing the transport of extracellular IPTG into the intracellular space. Furthermore, from (AVI.7) it holds that:

$$k_t = \frac{1}{V \cdot N_A} \cdot \frac{A \cdot u_m \cdot f_t}{V_{ex}} \cdot \frac{\left< I_{ex} \right> \cdot V_{ex} \cdot N_A}{\left< I_{ex} \right>} = \frac{u_m \cdot f_t \cdot A}{V_{E.coli}} = \frac{h_t \cdot A}{V_{E.coli}} \quad (AVI.9)$$

and this equation defines the permeability constant $h_t$ which is not a function of the cell size contrary to $k_t$. 
The derivation of the stochastic kinetic constant for the opposite transport
(intracellular IPTG out to the intracellular space) follows from the above.

Finally, for the case of the facilitated IPTG transport, one will have to take into
account that the LacY permeases are uniformly distributed on the membrane of the cell
covering a fraction $\phi_Y$ of the total membrane area, thus the collision volume will be:

$$
\delta V_{\text{coll}} = \delta A \cdot \phi_Y \cdot \left( -\nu_{\text{IPTG}} \cdot \mathbf{n}_A \right) \cdot \mathbf{1}_{\left( -\nu_{\text{IPTG}} \cdot \mathbf{n}_A > \phi \right)} \cdot \delta t
$$

(\text{AVI.10})

$\phi_Y$ can be mapped to the number of the LacY permeases existing in the cell and this is
how the propensity function $k_p \cdot [I_{\text{ex}}] \cdot Y$ is constructed.
References


Carrier, T., K. L. Jones and J. D. Keasling (2000). "mRNA stability and plasmid copy number effects on gene expression from an inducible promoter system." Biotechnology and Bioengineering 59: 666-672.


Fredrickson (1976). "Formulation of Structured Growth Models." *Biotechnology and*
Bioengineering XVIII: 1481-1486.


757-764.


